

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188	
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.				
1. AGENCY USE ONLY (Leave blank)		2. REPORT DATE 9.Jan.04		3. REPORT TYPE AND DATES COVERED DISSERTATION
4. TITLE AND SUBTITLE "EFFECTIVE MESOSCALE, SHORT-RANGE ENSEMBLE FORECASTING"			5. FUNDING NUMBERS	
6. AUTHOR(S) MAJ ECKEL FREDERICK A				
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) UNIVERSITY OF WASHINGTON			8. PERFORMING ORGANIZATION REPORT NUMBER  CI04-5	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) THE DEPARTMENT OF THE AIR FORCE AFIT/CIA, BLDG 125 2950 P STREET WPAFB OH 45433			10. SPONSORING/MONITORING AGENCY REPORT NUMBER	
11. SUPPLEMENTARY NOTES				
12a. DISTRIBUTION AVAILABILITY STATEMENT Unlimited distribution In Accordance With AFI 35-205/AFIT Sup			12b. DISTRIBUTION CODE	
<div style="border: 1px solid black; padding: 5px; text-align: center;"> <b>DISTRIBUTION STATEMENT A</b>            Approved for Public Release            Distribution Unlimited         </div>				
13. ABSTRACT (Maximum 200 words)				
20040121 089				
14. SUBJECT TERMS			15. NUMBER OF PAGES 224	
			16. PRICE CODE	
17. SECURITY CLASSIFICATION OF REPORT		18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT

# **Effective Mesoscale, Short-Range Ensemble Forecasting**

Frederick Anthony Eckel

A dissertation submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

University of Washington

2003

Program Authorized to Offer Degree: Department of Atmospheric Sciences

**DISTRIBUTION STATEMENT A**  
Approved for Public Release  
Distribution Unlimited

University of Washington  
Graduate School

This is to certify that I have examined this copy of a doctoral dissertation by

Frederick Anthony Eckel

and have found that it is complete and satisfactory in all respects,  
and that any and all revisions required by the final  
examining committee have been made.

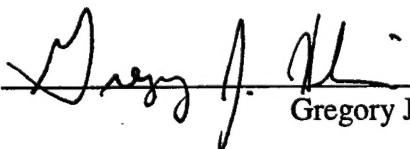
Chair of Supervisory Committee:

  
Clifford F. Mass

Reading Committee:

  
Clifford F. Mass

  
Dale R. Durran

  
Gregory J. Hakim

Date: 1 December 2003

In presenting this dissertation in partial fulfillment of the requirements for the Doctoral degree at the University of Washington, I agree that the Library shall make its copies freely available for inspection. I further agree that extensive copying of the dissertation is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Requests for copying or reproduction of this dissertation may be referred to Proquest Information and Learning, 30 North Zeeb Road, Ann Arbor, MI 48106-1346, to whom the author has granted "the right to reproduce and sell (a) copies of the manuscript in microform and/or (b) printed copies of the manuscript made from microform."

Signature Frederick A. Schul  
Date 1 December 2003

The views expressed in this dissertation are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.



University of Washington

**Abstract**

Effective Mesoscale, Short-Range Ensemble Forecasting

Frederick Anthony Eckel

Chair of Supervisory Committee:  
Professor Clifford F. Mass  
Department of Atmospheric Sciences

This study developed and evaluated a short-range ensemble forecasting (SREF) system with the goal of producing useful forecast probability (*FP*). Real-time, 0 to 48-h forecasts from four different SREF systems were compared for 129 forecast cases over the Pacific Northwest. Eight analyses from different operational forecast centers were used as initial conditions (ICs) for running the Fifth-Generation Pennsylvania State University–National Center of Atmospheric Research Mesoscale Model (MM5). Additional ICs were generated through linear combinations of the original 8 analyses, but this did not result in an increase in *FP* skill commensurate with the increase in ensemble size. It was also found that an ensemble made up of unequally likely members can be skillful as long as all members at least occasionally perform well.

Model error is a large source of forecast uncertainty and must be accounted for to maximize SREF utility, particularly for mesoscale, sensible weather phenomena. Inclusion of model perturbations in a SREF increased dispersion toward statistical consistency, but low dispersion remained problematic. Additionally, model perturbations notably improved *FP* skill (both reliability and resolution), revealing the significant influence of model uncertainty. Systematic model errors (i.e., biases) should always be removed from a SREF since they are a large part of forecast error but do not contribute to forecast uncertainty. A grid-based, 2-week, running-mean bias correction was shown to improve *FP* skill through: 1) better reliability by adjusting the ensemble mean toward the verification's mean, and 2) better resolution by reducing unrealistic ensemble variance.

Comparing the multimodel (each member uses a unique model) and the perturbed-model (each member uses a unique version of MM5) approaches for accounting for model uncertainty, it was found that a multimodel SREF exhibited greater dispersion (from more complete representation of model uncertainty) and superior performance. It was also found that smaller grid spacing leads to greater ensemble spread as smaller scales of motion are modeled. This study indicates substantial utility in current SREF systems and suggests several avenues for further improvement.

## Table of Contents

List of Figures.....	iii
List of Tables .....	vi
Acronyms .....	vii
Symbols .....	viii
Glossary .....	ix
Introduction .....	1
I. Background .....	5
A. EF Goal.....	5
B. The Requirements of EF.....	10
1. Predictability Error Growth .....	13
2. Representation of Analysis Uncertainty.....	19
a) Methods for Analysis Uncertainty Representation.....	21
3. Representation of Model Uncertainty .....	26
a) Perturbed-Model Theory .....	29
b) Research into EF and Model Uncertainty.....	31
4. Sufficient Ensemble Size .....	38
II. Methodology .....	52
A. Analysis Uncertainty .....	53
1. IC Strengths .....	59
2. IC Deficiencies.....	61
B. Model Uncertainty.....	65
1. Perturbed-Model Application .....	66

a) Model Options .....	69
b) Perturbations to Surface Boundary Parameters .....	70
2. Multimodel Application .....	75
C. Postprocessing and Analysis .....	76
1. Verification .....	76
2. Bias Correction .....	78
3. Forecast Probability Calculation .....	85
III. Results .....	123
A. Impact of Bias Correction .....	124
B. Model Uncertainty .....	128
1. Multimodel vs. Perturbed-model .....	128
a) Dispersion .....	128
b) Skill and Utility .....	133
2. Mesoscale: ACME <sup>core</sup> vs. ACME <sup>core+</sup> .....	135
a) Dispersion .....	136
b) Skill and Utility .....	140
C. ACME and Analysis Uncertainty .....	142
1. Skill and Utility .....	143
2. Dispersion .....	146
D. Future Research .....	148
IV. Summary .....	175
Appendix I: EF Statistical Toolbox .....	185
Appendix II: ACME <sup>core+</sup> Reference Data .....	201
References .....	219

## List of Figures

Figure 1. Example analysis, forecast, and climate PDFs for <i>MSLP</i> .....	44
Figure 2. Simplified depiction of EF .....	44
Figure 3. Error variance diagram examples.....	45
Figure 4. Dispersion diagram for a hypothetical 48-h forecasts.....	46
Figure 5. Simplified 2D generation of ICs from a best guess analysis.....	46
Figure 6. Tree diagram for an ensemble with 4 initial conditions and 3 model perturbations. ....	47
Figure 7. Impact of including model perturbations in a MREF.....	48
Figure 8. Sampling distributions of the sample mean and sample variance of an $N(0,1)$ PDF ....	49
Figure 9. Standard errors of the sampling distributions for increasing sample size .....	50
Figure 10. The 129 forecast case days of the research dataset over the 2002-2003 cool season..	91
Figure 11. Grid domains (Lambert conformal projections) of the SREF systems .....	91
Figure 12. Display of the mirroring of <i>RH</i> for three different values of the centroid <i>RH</i> .....	92
Figure 13. 2-D demonstration of the mirroring technique.....	92
Figure 14. Simulated EF attempts to represent a hypothetical forecast PDF .....	93
Figure 15. Sampling distribution of $\bar{x}$ using 5000 samples of size $n = 8$ .....	94
Figure 16. Sampling distribution of $s^2$ from 5000 samples of size $n = 8$ .....	94
Figure 17. Impact on forecast probability of the standard error in mean and variance .....	95
Figure 18. Mean absolute error from a single 48-h <i>MSLP</i> forecast of the eta member .....	96
Figure 19. Section of the 36-km resolution grid domain showing MM5 land use number .....	97
Figure 20. Sample surface boundary parameter PDFs .....	98
Figure 21. Example SST fields from 8 Jan 2003.....	99
Figure 22. Plot of the SST perturbations for member plus01 .....	100
Figure 23. Sample ACME <sup>core</sup> avn 3-h forecast .....	101

Figure 24. Scatter plots of 36-h forecast <i>MSLP</i> vs. centroid-analysis verification.....	102
Figure 25. <i>MSLP</i> bias for the avn member of ACME <sup>core</sup> .....	103
Figure 26. Scatter plots of 24-h forecast $T_2$ vs. RUC20 verification .....	104
Figure 27. Scatter plots of 39-h forecast $WS_{10}$ vs. RUC20 verification.....	105
Figure 28. Results of <i>MSLP</i> bias correction for PME .....	106
Figure 29. As in Figure 28 but for ACME <sup>core</sup> .....	107
Figure 30. As in Figure 28 but for ACME <sup>core+</sup> .....	108
Figure 31. As in Figure 28 but for ACME <sup>core</sup> $T_2$ data .....	109
Figure 32. As in Figure 28 but for ACME <sup>core+</sup> $T_2$ data.....	110
Figure 33. As in Figure 28 but for ACME <sup>core</sup> $WS_{10}$ data.....	111
Figure 34. As in Figure 28 but for ACME <sup>core+</sup> $WS_{10}$ data .....	112
Figure 35. Observation-based verification results of <i>MSLP</i> bias correction for ACME <sup>core</sup> .....	113
Figure 36. As in Figure 35 but for $T_2$ data .....	114
Figure 37. Example $WS_{10}$ PDF at a model grid point .....	115
Figure 38. Schematic calculation of <i>FP</i> by DV and UR .....	115
Figure 39. Schematic calculation of <i>FP</i> for an event threshold in the extreme right rank .....	116
Figure 40. Effect of ensemble size and <i>FP</i> calculation methodology on <i>FP</i> skill.....	117
Figure 41. Mean $Z_{500}$ and RMS of the time-filtered $Z_{500}$ .....	152
Figure 42. Reliability diagram for 36-h <i>FP</i> of the event $MSLP < 1001.0$ mb.....	153
Figure 43. <i>BSS</i> and its components for <i>FP</i> of the event $MSLP < 1001.0$ mb.....	154
Figure 44. <i>BSS</i> and its components for <i>FP</i> of $T_2 < 0^\circ\text{C}$ .....	155
Figure 45. Dispersion diagram for <i>MSLP</i> on the outer 36-km domain .....	156
Figure 46. Effect of bias correction on VRHs of forecast <i>MSLP</i> with a 24-h lead time .....	157
Figure 47. Dispersion diagrams for bias-corrected forecasts .....	158

Figure 48. VRHs that show the impact of including model diversity .....	160
Figure 49. Special dispersion diagrams. ....	161
Figure 50. $Z_{500}$ ensemble mean and $V_z$ for a *ACME <sup>core+</sup> and *PME forecast .....	162
Figure 51. <i>BSS</i> and its components for <i>FP</i> of <i>MSLP</i> < 1011 mb using ocean-masked data. ....	164
Figure 52. <i>BSS</i> and its components for <i>FP</i> of $T_2 < 0^\circ\text{C}$ using ocean-masked data.....	165
Figure 53. <i>BSS</i> improvement by *ACME <sup>core+</sup> over *ACME <sup>core</sup> .....	166
Figure 54. <i>BSS</i> and its components for <i>FP</i> of $WS_{10} > 18$ kt using ocean-masked data. ....	167
Figure 55. <i>ROCSS</i> for <i>FP</i> of <i>MSLP</i> < 1001 mb, $WS_{10} > 18$ kt, and $T_2 < 0^\circ\text{C}$ .....	168
Figure 56. Deterministic skill comparison of ACME members .....	169
Figure 57. <i>BSS</i> for $P(\textit{MSLP} < 1001 \text{ mb})$ for 8-member vs. 7-member SREF systems .....	170
Figure 58. <i>BSS</i> for $P(\textit{MSLP} < 1001 \text{ mb})$ for *ACME <sup>core</sup> and *ACME.....	171
Figure 59. <i>MSLP</i> dispersion diagram for *ACME <sup>core</sup> and *ACME.....	171
Figure 60. VRH comparisons between *ACME and *ACME <sup>core</sup> .....	172
Figure 61. $Z_{500}$ verification rank and $V_z$ for 36-h lead time of a forecast case .....	173
Figure 62. Example VRH for 1,781,676 trials of $T_{850}$ using bias-corrected ACME <sup>core</sup> .....	196
Figure 63. Hypothetical PDF and for an idealized EF case and realistic EF case .....	196
Figure 64. Reliability diagram for data in Table 10 .....	197
Figure 65. Graph all possible values of the uncertainty term in the <i>BS</i> .....	197
Figure 66. Contingency table of signal detection theory. ....	198
Figure 67. ROC for the data in Table 11 .....	198
Figure 68. Plots of the SST perturbations for each member of ACME <sup>core+</sup> .....	205

## List of Tables

Table 1. Abridged list of three categorical sources of model error .....	51
Table 2. Brief description of the four SREF systems .....	118
Table 3. The eight analysis/forecast modeling systems of the PME .....	119
Table 4. List of ACME <sup>core+</sup> model versions .....	120
Table 5. Data of three core samples of $n = 8$ .....	121
Table 6. The standard MM5 land use table .....	122
Table 7. BSS data for $MSLP < 1001$ mb .....	174
Table 8. Skill score comparison between *ACME and *ACME <sup>core</sup> at the 36-h lead time .....	174
Table 9. Two sets of hypothetical EFs of $T_{850}$ ordered from least to greatest.....	199
Table 10. Summary of 22,402 probability forecasts of 24-h cumulative precip. > 0.25 inch ....	200
Table 11. Calculated values for the ROC for the same source data as in Table 10.....	200
Table 12. Gamma variables for the 48 PDFs used to generate albedo values .....	209
Table 13. Gamma variables for the 48 PDFs used to generate moisture availability values.....	209
Table 14. Gamma variables for the 48 PDFs used to generate roughness length values .....	210
Table 15. MM5 land use tables used for ACME <sup>core+</sup> .....	211

## Acronyms

ACME – analysis-centroid mirroring ensemble  
avn – aviation model  
BGM – breeding of growing modes  
CDF – cumulative density function  
CMC –  
DV – democratic voting  
EF – ensemble forecast  
ECMWF – European Centre for Medium-Range Weather Forecasts  
EnKF – ensemble Kalman filter  
FNMOC – Fleet Numerical Meteorology and Oceanography Center  
GEM – Global Environmental Multiscale model  
GFS – Global Forecast System  
GPH – geopotential height  
IC – initial condition  
LAM – limited area model  
LBC – lateral boundary condition  
LSM – land surface model  
MM5 – fifth-generation mesoscale model  
MMMA – multimodel multianalysis  
MREF – medium-range ensemble forecast  
NCEP – National Center for Environmental Prediction  
NWP – numerical weather prediction  
OI – optimal interpolation  
OTIS – Optimum Thermal Interpolation System  
PBL – planetary boundary layer  
PDF – probability density function  
PME – poor man's ensemble  
PMMA – perturbed-model multianalysis ensemble  
ROC – relative operating characteristic  
RUC20 – Rapid Update Cycle 20-km resolution modeling system  
PQPF – probabilistic quantitative precipitation forecast  
SBP – surface boundary parameter  
SMMA – single model multianalysis  
SSE – system simulation experiment  
SST – sea surface temperature  
SREF – short-range ensemble forecast  
SV – singular vector  
UKMO – United Kingdom Meteorological Office  
UR – uniform ranks  
VRH – verification rank histogram



## Symbols

This is a list of symbols used repeatedly but not a complete list of all symbols used.

*BSS* – Brier skill score

$\bar{e}$  – ensemble mean

$\bar{E}_0$  – analysis error

$\hat{\bar{E}}_0$  – estimated analysis error

*FP* – forecast probability

*MAE* – mean absolute error

*MSE* – mean-square error

*MSLP* – mean sea level pressure (mb)

*n* – ensemble or sample size

*ORF* – observed relative frequency

*rel* – reliability

*res* – resolution

*RH* – relative humidity (%)

*RMSE* – root-mean-square error

*ROCSS* – relative operating characteristic skill score

*SC* – sample climatology (%)

*s* – sample standard deviation

*T*<sub>2</sub> – 2-m temperature (°C)

*unc* – uncertainty

*VOP* – verification outlier percentage (%)

*V* – verification

*V*<sub>Z</sub> – standardized verification

*WS*<sub>10</sub> – 10-m wind speed (m s<sup>-1</sup>)

$\bar{x}$  – sample mean

*Z*<sub>500</sub> – 500-mb geopotential height (gpm)

$\sigma$  – population standard deviation

$\sigma^2$  – population variance

$\sigma_c^2$  – climatic variance

$\tau$  – event threshold

$\mu$  – population mean

$\mu_c$  – climatic mean

## Glossary

**Analysis-Centroid Mirroring Ensemble (ACME)** – The primary EF of this research whose ICs are made up of various independent analyses and their mirrors in model phase space using the mean analysis, or centroid, as the reflection point.

**Analysis** – A complete description of the state of the atmosphere normally derived from some combination of a first guess (from an NWP model) and observations. It defines values for all state variables at all model grid points. See also *initial condition*.

**Analysis PDF** – A probability density function of possible atmospheric states from which an analysis (or IC) is a random sample. It is defined by a set of ICs.

**Attractor** – The union in phase space of all naturally occurring states of a dynamical system.

**Centroid** – The MM5 forecast that used the centroid analysis as its IC.

**Centroid Analysis** – The mean of many different analyses produced by operational forecast centers.

**Climate PDF** – A probability density function of all possible atmospheric states for a given time of year.

**Dispersion** – This term is normally used interchangeably with predictability error growth but we are using a modified definition. Dispersion is the increase in ensemble spread from the spread in the ICs.

**Ensemble Forecast (EF)** – A collection of many different, equally likely NWP model solutions derived from various ICs and/or models. Its purpose is to build a forecast PDF from which forecast uncertainty and probabilistic forecasts can be derived.

**Ensemble Spread** – The unbiased (divided by  $n - 1$  instead of  $n$ ) variance of the ensemble members.

**Encompass Truth** – When the verification value is bound by ensemble members (i.e., lowest forecast in the ensemble < verification value < highest forecast in the ensemble).

**Ensemble Mean** – The average of all ensemble members at a certain forecast lead time.

**Ensemble Member** – One of the many individual forecast model runs that make up the entire ensemble.

**Event** – The occurrence above or below a threshold value (i.e., *event threshold*) of some parameter, either instantaneously or over a period of time (e.g., surface temperature less than freezing at 12Z; or 12-h cumulative precipitation greater than 0.5 in; or wind speed above 20 kt).

**Event Threshold** – The critical value of a parameter for an *event*.

**Forecast Event** – see *event*

**Forecast Lead Time** – The amount of time (usually in hours) from the initialization time of a forecast cycle to an instant being forecast.

**Forecast PDF** – A probability density function of possible future atmospheric states, defined by the entire collection of ensemble members.

**FP, Forecast Probability** – The predicted chance of occurrence of a parameter exceeding some threshold. (EX: 35% chance of cumulative precipitation greater than 10 mm in three hours)

**Initial Condition (IC)** – A starting point in a NWP model's phase space required to run the model. Note that an analysis is always considered an IC but an IC may not be an analysis since it may be generated as a perturbation to an analysis.

**Ideal Ensemble** – An ensemble that completely represents all uncertainty so that the true state is always a random draw from the EF's estimated forecast PDF.

**Lateral Boundary Condition (LBC)** – The state variables periodically updated (normally by model data on a larger domain) on the domain edges of a limited area model.

**Member** – see *ensemble member*

**Monte Carlo** – The method of generating ICs by adding scaled, random noise to the best guess analysis.

**Multimodel Multianalysis Ensemble (MMMA)** – An EF approach designed to include representation of both analysis uncertainty and model uncertainty, accomplished by applying a set of ICs to different NWP models.

**Numerical Weather Prediction (NWP)** – The mimicking of the evolution of the atmosphere by modeling the time rate of change of the state variables with approximations of the governing laws of fluid dynamics, momentum, gas, and entropy over a discrete domain.

**Observed Relative Frequency (ORF)** – For a bin of *FP* (i.e., a set of forecasts with similar *FP* values), *ORF* is the number of occurrences observed above the event threshold divided by the number of forecasts in the bin.

**Probability Density Function (PDF)** – (Devore, 1995) A function  $f(x)$  such that for any two numbers  $a$  and  $b$  with  $a \leq b$ , the probability  $P$  that a continuous random variable  $X$  takes on a value between  $a$  and  $b$  is:

$$P(a \leq X \leq b) = \int_a^b f(x) dx$$

**Phase Space** – A multi-dimensional plotting region where all time dependent variables of a dynamical system are represented by a unique dimension. The instantaneous state of the system is then completely described by a single point, and the system's evolution is a line or *trajectory*.

**Physics Parameterization** – An atmospheric quantity, factor, or process which is not completely known and/or of too small a scale to be properly represented at a given resolution in a NWP model.

**Perturbed-Model Multianalysis Ensemble (PMMA)** – An EF approach designed to include representation of both analysis uncertainty and model uncertainty, accomplished by applying a set of ICs to different (perturbed) versions of the same basic NWP model.

**Poor Man's Ensemble (PME)** – The EF comprised of the model runs from different operational centers. The "poor man" refers to the fact that the only cost involved is downloading and organizing the data.

**Portray Truth** – When the verification occurs within three standard deviations from the mean of the forecast PDF of an EF. A verification value that is not portrayed is therefore an outlier with respect to that PDF. Note that truth may be portrayed by not necessarily encompassed.

**Predictability Error Growth** – The magnitude of the difference between a forecast solution and the verification as the forecast diverges from truth with increasing forecast lead time. The rate at which errors grow determines the point at which the errors become saturated (i.e., equal to the average error of the climatic mean) and predictability is lost.

**Reliability** – The ability of FP to match the ORF.

**Resolution** – The ability of an ensemble system to distinguish between events and non-events. The sharpness of an ensemble's forecast PDFs.

**Spread** – see *ensemble spread*

**Short-Range Ensemble Forecast (SREF)** – An EF designed to build a forecast PDF for short range (normally 0-48 h but can be up to 60 h), mesoscale weather phenomena.

**System Simulation Experiment (SSE)** – A method to isolate and diagnose error sources in an NWP model by running parallel model integrations with slightly different versions of the model in each run.

**State Variables** – The basic set of meteorological parameters required to describe the atmosphere at a single point. E.g., horizontal and vertical wind components ( $u$ ,  $v$ ,  $w$ ), temperature ( $T$ ), moisture ( $q$ ), and pressure ( $p$ ) or geopotential height ( $\Phi$ ).

**Statistical Consistency** – The ability of the mean square error of the ensemble mean to match the average ensemble variance over a large sample of data. The requirement that the verification be a random sample from the PDF of the EF.

**Surface Boundary Parameter (SBP)** – A spatially dependent variable that affects the atmosphere's evolution and is estimated in an NWP model (Example: sea surface temperature)

**Target Variance** – The total amount of uncertainty (i.e., variance) that should be produced by a well-tuned EF system in order to achieve statistical consistency. The mean-square error in the spatially and temporally averaged EF mean.

**Verification** – The observed value of an atmospheric parameter at a specific forecast lead time used to verify a forecast of that parameter.

**Verification Rank Histogram (VRH)** – A tool for evaluating an EF made from repeatedly tallying the rank of the verification when pooled with the ordered forecast values from an EF.

## Acknowledgments

I must begin by thanking the United States Air Force and Colonel Nathan Feldman for giving me this opportunity to pursue an advanced degree while continuing to serve on active duty. Additionally, this research was supported by the Department of Defense Multidisciplinary University Research Initiative (MURI) program administered by the Office of Naval Research under Grant N00014-01-10745, as well as the National Weather Service.

I would like to thank the following people for their contribution to this research:

First, big thanks to my advisor Cliff Mass who besides providing expert scientific advice, supplied the extensive resources necessary for this research. I also appreciated how he gave me plenty of flexibility in exploring my ideas but also kept me well grounded and on track to meet my graduation goal. Thanks also to all my committee members, Dale Durran, Greg Hakim, Brad Colman, and Dave Baumhefner, for your outstanding review of this dissertation and extremely helpful inputs.

Over the past three years Eric Grit and I worked together to make tremendous progress on short-range ensemble forecasting—much more than the sum of what we would have accomplished separately. I would like to thank him for all of our lengthy, enlightening discussions and for his efforts and computer system expertise that made our SREF systems work so well in real time. Thanks also to Dave Ovens for all his help in designing the complex computer processing of our systems.

I owe many thanks to my stepmother [REDACTED] who did a fantastic, professional job editing this dissertation. Lastly, thanks to all my classmates and friends who made my time at UW unforgettably fun.

## Introduction

This dissertation describes a research study in the field of ensemble forecasting (EF), accomplished at the University of Washington's Atmospheric Sciences Department under the supervision of Dr. Clifford Mass. As opposed to the more common, deterministic-style numerical weather prediction (NWP) where only a single model run is considered, EF is stochastic in nature, using multiple runs of an NWP model with slightly different initial conditions (ICs) and/or model variations. The resulting set of solutions defines a probabilistic distribution of future states of the atmosphere based on the inherent uncertainties in the analysis and/or in the model.

When Vilhelm Bjerknes laid the groundwork for NWP in the early 20<sup>th</sup> century, he noted that errors in the prognosis would arise from both an inaccurate IC and a deficient model (Bjerknes et al., 1911). We can only speculate whether Bjerknes realized that there is a major difference between the character of these two error sources, or whether he believed that both problems could eventually be reduced to insignificance. Analysis error is the predominant contributor to the nonlinear error growth that limits predictability (Lorenz, 1969; Leith, 1974). It may be possible to create a nearly perfect model but even with a nearly perfect analysis, IC errors will grow far beyond the model error.

Long before the advent of NWP, Jules Henri Poincare, a contemporary to Bjerknes, explained the differences between the error sources with profound clarity (Poincare, 1914):

"If we knew exactly the laws of nature and the situation of the universe at the initial moment, we could predict exactly the situation of that same universe at a succeeding moment. But even if it were the case that the natural laws had no longer any secret for us, we could still only know the initial situation *approximately*. If that enabled us to predict the succeeding situation *with the same approximation*, that is all we require, and we should say that the phenomenon had been predicted, that it is governed by laws. But it is not always so; it may happen that small differences in the initial conditions produce very great ones in the final phenomena. A small error in the former will produce an enormous error in the latter. Prediction becomes impossible, and we have the fortuitous phenomenon."

The “fortuitous phenomenon” is one that appears to behave by chance but is actually governed by deterministic laws, and “about which the calculation of probabilities will give us provisional information.” Poincare did not present this as a purely philosophical idea, but gave concrete examples of its application, including meteorology:

“The meteorologist sees very well that the equilibrium is unstable, that a cyclone will be formed somewhere, but exactly where they are not in a position to say; a tenth of a degree more or less at any given point, and the cyclone will burst here and not there, and extend its ravages over districts it would otherwise have spared. If they had been aware of this tenth of a degree, they could have known it beforehand, but the observations were neither sufficiently comprehensive nor sufficiently precise, and that is the reason why it all seems due to the intervention of chance.”

The potential for small IC error to produce large forecast errors and the value of a probabilistic forecast are exactly what EF is all about. The only complete way to make a prediction of the future state of the atmosphere is to include the inherent uncertainty as part of the forecast process.

Unfortunately, the science of meteorology was too primitive at this point to apply Poincare’s premise. The significance was lost until after the development of deterministic NWP through the efforts of Lewis Richardson (the first to solve the atmospheric primitive equations with numerical methods), Carl-Gustaf Rossby (developed simplified dynamics capable of producing an adequate analysis), John von Neumann (applied NWP to computers), and Jule Charney (developed the filtered equations for the first successful NWP forecast) (Lorenz, 1993). Without the contributions of these scientists, NWP (and therefore EF) would not be at the highly developed state that it is at today.

Edward Lorenz (1963) rediscovered the ideas of Poincare and brought to light their impact on NWP. He demonstrated that the atmosphere is a chaotic dynamical system and that even if you could create a perfect model, predictability is limited by sensitivity to the imprecise ICs. This explained the primary reason for the limitations of deterministic NWP, which by this time was meeting with some success.



Epstein (1969) realized that this sensitivity to ICs made deterministic NWP an inadequate method for atmospheric prediction. In response, he formulated a stochastic dynamic forecast model designed to directly forecast the mean and variance of state variables (rather than simply a single deterministic value with unknown error) by incorporating uncertainty into the prognostic equations. This is a more comprehensive way to consider the future state of the atmosphere, but it requires overwhelming computational power. Leith (1974) proposed the method of EF as an approximation to stochastic dynamic forecasting, focusing primarily on IC error or what he termed "internal error." Unfortunately, this method for probabilistic forecasting was impractical at that time since there was only enough computer power to run an NWP model once and not the multiple runs proposed by Leith.

By the 1990s, increasing computer power allowed application of Leith's ensemble method for dealing with the forecast problem raised by Poincare. Successful medium-range (2 – 10 days) ensemble forecasting (MREF) began at the National Centers for Environmental Prediction (NCEP) and the European Centre for Medium-Range Weather Forecasts (ECMWF) (Toth and Kalnay, 1993; Tracton and Kalnay, 1993; Molteni et al., 1996). Operational use of short-range (0 – 48 h) ensemble forecasting (SREF) has lagged behind because compared to MREF, it has proven to be more difficult to design a SREF that can consistently capture all or at least most of the short-range forecast uncertainty. The potential benefits of SREF have not yet been fully realized and the value of SREF remains an open question (Hamill et al., 2000a). Some of the reasons for the difficulty of SREF compared to MREF, which will be addressed throughout this dissertation, may be:

1. The smaller-scale, surface parameters of interest in the short-range are less predictable so their errors may saturate too quickly for an ensemble to be of use.

2. Model uncertainty likely has a more significant impact on small-scale, surface parameters and is difficult to include in an ensemble since model errors are poorly understood.
3. The best method for defining the ICs for SREF is unclear since error growth is primarily linear in the short-range. For MREF, a wide variety of methods for defining ICs have proven useful since nonlinear error growth in the medium range allows any IC differences to grow to represent a large spread of solutions.

The goal of this research is to evaluate and find ways to improve the value of SREF by applying ensemble methods to short-range, mesoscale, atmospheric modeling for forecasting of sensible weather at the surface (e.g., surface temperature and surface wind). For this research a SREF system was built that ran the Fifth-Generation Pennsylvania State University–National Center of Atmospheric Research Mesoscale Model (MM5) using analyses from different operational forecast centers as ensemble ICs. This was not an attempt to build an ideal SREF but rather an opportunity to realize most of the potential SREF benefits by employing sub-optimal but sound methods that are currently computationally feasible. With such a system, we were able to address basic SREF issues that will apply to the development of more optimal SREF systems of the future.

Chapter I covers background material of EF and SREF. Chapter II discusses the methodology of the techniques, ideas, and procedures involved in this research. Chapter III details the results and findings. Chapter IV provides a summary of the entire research project. The appendices provide important reference material and technical information. Additionally, a glossary, list of acronyms, and a list of symbols are included for quick reference.

## I. Background

### A. EF Goal

The fundamental goal of EF is to produce a forecast probability density function (PDF) of possible future states of the atmosphere from which the true state is consistently a random sample (Talagrand et al., 1999). Upon reaching this goal, there are three general applications of EF (Epstein 1969; Leith 1974):

- 1) Use the EF mean to improve deterministic forecast skill and maximize predictability.
- 2) Predict forecast skill using the EF spread.
- 3) Predict the probability of future weather events.

Of these three applications, this research will focus primarily on the third since it is the key for making dramatic improvements in the value of weather forecasting. (In fact, we will purposely avoid the second application, the relationship between spread and skill, since fellow graduate student Eric Gritmit is investigating that using the same data.) Essential to understanding the EF goal and these applications is a clear distinction between four different theoretical notions: the true state of the atmosphere, the analysis PDF, the forecast PDF, and the climate PDF.

The true state is a vector ( $\vec{T}$ ) of state variables having infinite dimension and infinite precision that completely describes the atmosphere at some instant. In other words, it is the exact value of temperature, pressure, humidity, etc. throughout the atmosphere. In terms of chaos theory,  $\vec{T}$  is a phase space vector lying somewhere on the atmosphere's *attractor*. (Following Lorenz (1993), the term attractor will herein be used as the union in phase space of all naturally occurring states of a dynamical system.)

The analysis PDF is a frequency distribution of possible concurrent states from which an analysis is a random sample. This PDF exists only as an abstraction, arising from our limited

capability to observe and analyze the atmosphere at any point in time. It is a cloud of states (Leith, 1974) within phase space that encompasses a small region about the atmosphere's attractor and is dense in the middle, slowly thinning outward. The size and shape of the cloud represents our uncertainty in the true state as well as in the atmosphere's attractor since much of the cloud may lie off the attractor. We may consider the true state to be a random sample from the analysis PDF, but that is just our illusion. The true state is a deterministic result of the laws of nature (Lorenz, 1993) and not a sample from our vague view of reality. In fact, the analysis PDF is totally defined by our analysis capability, or lack thereof. The better and more complete our objective analysis process, the narrower (i.e., less uncertain) the analysis PDF.

The forecast PDF is similar to the analysis PDF except that a random sample from it is a possible future state, rather than a possible current state. The forecast PDF is a frequency distribution that represents our uncertainty in the prediction of the true state. It is also a cloud in phase space floating about a small region of the atmosphere's attractor, but it is naturally larger and more diffuse than the analysis' cloud since the atmosphere is a chaotic system. A NWP forecast evolves from an analysis so the forecast PDF is defined by both our analysis and forecast capability, or lack thereof. The better our analysis and NWP model, the narrower the forecast PDF will be at any forecast lead time.

Lastly, the climate PDF is a frequency distribution of all possible states of the atmosphere, or the possible states for one season. It can be thought of as a forecast PDF from an *ideal ensemble* with a very long lead time. The term "ideal ensemble" will be used to mean an ensemble that completely represents all uncertainty so that the true state is always a random draw from the EF's estimated forecast PDF. An extended run (on the order of weeks) with many members from such an ensemble will produce the season's climate PDF since the ensemble members will spread out to cover the full spectrum of climatologically possible states. Unlike the analysis and forecast

PDFs that change form based on the skill of atmospheric observation and modeling, the climate PDF is more concrete and defined by the variability of nature.

While it is impossible to visualize these PDFs completely because of their extremely large number of degrees of freedom, we can view limited slices (single variable over limited region) to demonstrate the characteristics described above. Figure 1a is a histogram of mean sea level pressure (*MSLP*) observations from the Aviation Model analysis over our research 36-km domain (see Figure 11) for one winter season. The distribution is obviously not Gaussian, and a good fit may be a Weibull PDF (Devore, 1995). The important point is that this distribution describes all possible values of truth, limited by the fact that it was derived by a model analysis. When we then create an analysis PDF (Figure 1b) to try to represent a value of *MSLP* at one point and one time, we cover a narrow region of the climate PDF. The forecast PDF does the same thing but must cover a wider region since it is more uncertain.

EF is often described as the process of sampling from the forecast PDF (Hamill, 2000), which is equivalent to imagining EF as the attempt to construct a good estimate of the forecast PDF. The actual forecast PDF can never be known since it would require an ideal ensemble of infinite size to produce it. In fact, the forecast PDF is often not well represented by EF due to limited sampling and inadequate representation of analysis and model uncertainty. Herein lies the genuine and often overlooked difficulty of EF. Not only do we have to deal with the fact that we see the future state as a PDF, but we also have significant uncertainty in that PDF. This uncertainty in our prediction of uncertainty is the real challenge to EF and has implications for EF verification as well. Just as we can never know the actual analysis error, we also can never know the actual error in an EF's estimate of the forecast PDF. This makes it extremely difficult to evaluate an EF because, when the true state is not encompassed, it is difficult to determine if that was a result of a bad forecast PDF or simply undersampling of a good PDF.

Figure 2 is a simplified demonstration of EF where the complex distribution of possible atmospheric states is represented by a two-dimensional, normal PDF. Alternatively, one can think of the displayed PDFs as a distribution of possible values for a single parameter (such as temperature) at a single location. In Figure 2a, an ideal ensemble correctly estimates both the analysis and forecast PDF, simulated by histogramming 500 random samples from the actual PDFs (gray solid curves). Figure 2b shows a typical ensemble with incorrect *location* (mean,  $\mu$ ) and *spread* (standard deviation of the ensemble,  $\sigma$ ) in its estimation of the analysis PDF, which then worsens in the forecast PDFs. An ensemble that errs in one or both of these quantities fails to realistically represent the actual uncertainty of where truth lies (Hamill, 2000). In other words, the true state can not be considered a random sample from the ensemble's estimate of the analysis or forecast PDFs. Possible causes of this failure will be discussed in the next section.

The long-term ability of an EF to correctly estimate the mean and spread of the forecast PDF can be revealed by verifying the ensemble mean (i.e., verification – EF mean) over a large sample of forecasts. A poor estimate of the mean of the forecast PDF is revealed by a significant bias in the EF mean's error. A problem in ensemble spread is found by comparing the magnitude of the EF mean's error with the ensemble spread, which should be comparable (Buizza, 1995; Talagrand, 1999; Hamill et al., 2000a). For most EF systems, it has been found that the error in the ensemble mean exceeds the ensemble spread, revealing a spread that is insufficient to consistently encompass truth.

The term "encompass truth" will be used to mean that the verification value is completely bound by the EF members. It should be clear that for an ideal ensemble with an infinite number of members, truth must be encompassed by the EF's forecast PDF. This behavior gets a bit vague when dealing with a finite number of members. Occasional failure to encompass truth is expected since truth can occur in the tail of the forecast PDF, beyond the most extreme ensemble

member. If truth were to fall outside the ensemble too often (above what is expected because of undersampling), the ensemble is obviously underdispersive. The problem is that simply considering how often truth is encompassed does not reveal when truth is an *outlier* with respect to the EF's approximate forecast PDF (i.e., when truth is not sampled from the same PDF as the EF members). We will therefore use the term "portray truth" to mean that the verification occurs within three standard deviations from the ensemble mean.

Figure 2 also displays how an ensemble PDF can be used to produce a *forecast probability (FP)* for some *forecast event*. We define a "forecast event" as the occurrence above or below a threshold value (called the *event threshold*) of some parameter, either instantaneously or over a period of time (e.g., temperature less than freezing, or 12-h precipitation greater than 0.5 in). For illustrative purposes, let's say the PDF random variable in Figure 2 is wind speed at some location and we want to know the chance of exceeding 20 kt. The ensemble-based *FP* is given by the area under the PDF to the right of the event threshold, known as the  $1-p$  value in statistics (shaded area in Figure 2).

An EF that consistently and accurately estimates the forecast PDF will display a high degree of *reliability* (i.e., the *FP* will match up with the *observed relative frequency (ORF)*, given a large number of forecast/observation data pairs). For example, consider 100 instances in which  $FP = 35\%$  chance of wind speed  $\geq 20$  kt. We should expect the wind speed to be faster than 20 kt in exactly any 35 of those instances, for an  $ORF = 35\%$ . It is clear for the deficient EF of Figure 2b that  $FP \neq ORF$ . However, this does not mean such an EF is useless since *FP* may still have valuable predictive skill without perfect reliability (see Appendix I).

The other component of *FP* skill is *resolution*, the ability to distinguish between events and non-events. Binary-type forecasts (i.e., yes, no or 100%, 0%) have the highest possible resolution (regardless of their reliability) since they maximize the distinction between when an event may or

may not happen. For fully probabilistic forecasts of any given event threshold, more certain forecast events (i.e., greater agreement among EF members) tend to have extreme *FP* (i.e., near 0% or 100%) and thus a higher resolution while less certain forecast events have midrange *FP* and lower resolution.

The utility (i.e., value to a user) of forecasts depends upon both their reliability and their resolution. Ensemble-based *FP* normally has lower resolution but greater utility compared to binary forecasts that suffer from poor reliability. The strength of ensemble-based *FP* comes from the fact that it combines all the information of EF into a single product that encapsulates the uncertainty in the forecast process. Indeed, *FP* is the icing on the cake for EF because for practical application, the overwhelming amount of data from multiple forecast solutions must be condensed.

A common misconception is that EF can extend the atmosphere's limit of predictability. However, the predictability limit is established primarily by the analysis error (Lorenz, 1969; Leith, 1974; Rabier et al., 1996; Errico et al., 2002). EF does not correct for this error but uses it as a basis to estimate the error growth during the forecast period. Therefore, EF can not extend predictability, but it can reveal the predictability limit. When comparing long-term error statistics of an ensemble mean vs. a deterministic forecast, it may appear that EF extends predictability simply because a deterministic forecast often, but not necessarily, has much greater error. (This concept will be explained further below.)

## **B. The Requirements of EF**

In this section we will discuss theoretical aspects of how an EF must be designed to account for the uncertainty of weather forecasting. This section will also describe how the challenge of EF has been met to date.



There are three basic requirements to meet in attempting to run a skillful EF system (Palmer et al., 1990):

- 1) **Representation of Analysis Uncertainty:** Ensemble ICs must be formulated such that differences between ICs represent analysis error and the true analysis is a random sample from the EF's analysis PDF.
- 2) **Representation of Model Uncertainty:** If model error is significant, the resulting uncertainty must be accounted for in the EF.
- 3) **Sufficient Ensemble Size:** There must be enough members in the ensemble to produce a thorough statistical sampling of the forecast PDF.

The level to which the three requirements for EF must be met generally depends upon the specific application. An important caveat with the first two requirements is that for an EF to have a chance at being effective, the portion of forecast error due to IC uncertainty must be larger than the portion due to model error (Murphy 1988; Palmer et al., 1990). If model uncertainty dominates then the EF's approximate forecast PDF may be of little value because its sample states would be so much different compared to the true atmosphere. Such a PDF would have to be very wide to portray the true state and thus would have extremely low resolution. This may be a reason for the difficulty of SREF since a mesoscale model is often deficient in representing the small-scale phenomena of interest. Research so far (Houtekamer et al., 1996; Buizza et al., 1999; Stensrud et al., 2000; Mylne et al., 2002) has shown that forecast errors due to the model are significant for EF, but the model's contribution to forecast error relative to the contribution from IC error has not been clearly demonstrated.

The two sources of uncertainty (analysis and model) present very different challenges for EF. Their dissimilarity may seem obvious but clarification from the point of view of chaos theory is

enlightening. It is a question of starting the ensemble in the correct spot in phase space versus evolving the solution on the correct attractor.

If a perfect model is assumed, the only concern is the analysis uncertainty. In this context, forecast errors of a single deterministic model run arise solely because of IC sensitivity in a dynamical system. To deal with this, a best guess analysis (generated by some objective analysis cycle) can be randomly perturbed (scaled by the magnitude of the typical analysis error)  $n - 1$  times, to produce a total of  $n$  ICs. This Gaussian cloud of ICs then represents the uncertainty in the true state at the initialization and defines the analysis PDF. Upon running the  $n$  ensemble members in the perfect model, the true state will be well portrayed, given a large  $n$ .

This scenario is the ideal EF system depicted in Figure 2a. Figure 2b shows what can happen when there are problems in producing the ICs. Deficiencies in the analysis cycle can shift the location of the EF's estimated analysis PDF and poorly scaled or formulated perturbations can affect the spread. So even with a perfect model, the EF produces poor estimated forecast PDFs.

Now assume the reverse condition of a perfect analysis but an erred model. In this context, forecast errors arise because our modeled solution evolves on an attractor that differs from the atmospheric attractor. To deal with this it is necessary to perturb about the uncertainty within the model, an even more complex issue than perturbing about analysis uncertainty. What is needed is  $n$  different, valid models representing the uncertainty in the atmospheric attractor. Each ensemble member then evolves on a unique but erred estimate of the true attractor, and truth would again be well portrayed for a large  $n$ .

Figure 2 can also be used to imagine the erred-model scenario, except the analysis PDF must be imagined as an infinite spike at one value (i.e., no analysis uncertainty). The forecast PDF still spreads out with increasing lead time as the members' solutions evolve on different attractors.

Figure 2b shows what happens when the model uncertainty is not well accounted for. Model bias

can shift the location of the EF's estimated forecast PDF and insufficient (excessive) representation of the model error causes spread to be too low (high). So even with a perfect analysis, the EF again produces poor estimated forecast PDFs.

The third requirement of EF, the need for a large ensemble size, adds a further twist to Figure 2. When it comes to implementing an EF system, computational resources constrain the system to a finite and often very limited number of ensemble members. This has severe implications for the EF's ability to consistently construct a reasonable forecast PDF. This effect is often overlooked when analyzing an EF system, with more attention being paid to the first two requirements. Consider once again the ideal EF of Figure 2a where  $n$  is large; then imagine taking a subset of that same EF with only  $n = 8$ . While it is still possible to produce the same PDFs, error is more likely in both the PDF location and spread—yet another independent way to produce Figure 2b.

A complete way to interpret an EF's lack of success, such as depicted in Figure 2b, is that it resulted from failure to meet all three EF requirements, thus making diagnosis of the source of an ensemble's problems very challenging. An ensemble's inability to produce an accurate estimate of the forecast PDF comes simultaneously from deficient accounting for analysis uncertainty, deficient accounting for model uncertainty, and incomplete sampling. Failure to adequately meet any of the three requirements leads to inaccurate depiction of *predictability error growth* by the ensemble. This is a fundamental concept in EF so we will explain it in detail first, before elaborating further on each of the three requirements separately.

### 1. Predictability Error Growth

One way to describe and understand predictability error growth is with an *error variance diagram* (Figure 3). This diagram, designed by David Baumhefner (2000) and based on the work of Leith (1974), is a visual display of the basic limitations and potential benefits of EF.

Predictability error growth is a measure of how forecast errors grow on average and when (at what lead time) predictability is lost. In Figure 3, predictability error growth is plotted using the spatially averaged variance of the forecast error over increasing lead time for a particular meteorological parameter from a single model run, called the control run, computed by:

$$(s_{f^*}^2)_t = \frac{1}{M} \sum_{m=1}^M \left( (f_{m,t} - o_{m,t}) - \bar{f}_t^* \right)^2 \quad (1)$$

where  $M$  is the number of forecast points,  $f_{m,t}$  is a single forecast at point  $m$  and lead time  $t$ ,  $o_{m,t}$  is the verifying observation, and  $\bar{f}_t^*$  is the average control forecast error for all  $M$  points at lead time  $t$ . (The asterisk is used to denote the error in a variable.) To simplify the explanations in this section, we are restricting the analysis to a single EF case (i.e., one forecast cycle) over a grid of  $M$  points, but the error variance diagram is normally applied to many EF cases. (I.e., the curves of Figure 3 are actually averages over many cases.)

When the error variance of the control reaches the climatic variance ( $\sigma_c^2$ , the long-term, spatially and temporally averaged variance of the parameter being forecast), the average error of the deterministic forecast is the same as the average error of the climatic mean ( $\mu_c$ ). This is the limit of predictability for the control forecast. For lead times beyond that point, the climatic mean is a better forecast. For a well-calibrated model (i.e., dispersive characteristics equivalent to nature) the curve asymptotes to twice the climatic variance ( $2\sigma_c^2$ ), a feature discussed in detail below.

The curve that is closely related to the control's error variance is the variance of the differences between ensemble members (IC perturbations only):

$$\langle s_d^2 \rangle_t = \frac{1}{M} \sum_{m=1}^M \left( \frac{1}{D} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n ((e_i - e_j) - \bar{d})^2 \right) \quad (2)$$

where  $n$  is the number of ensemble members and  $e_i$  is an ensemble member forecast at a particular point and lead time  $t$ . (Notice that the difference between forecasts is signed.) The brackets,  $\langle \rangle$ , denote an average of all forecast points.  $D$  is the number of differences among the members:

$$D = 2 \sum_{i=1}^{n-1} (n-i) = n(n-1) \quad (3)$$

and  $\bar{d}$  is the mean of the set of  $D$  differences:

$$\bar{d} = \frac{1}{D} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n (e_i - e_j) \quad (4)$$

The differences curve reveals how the ensemble members diverge with increasing lead time and depends upon the intrinsic variance of the model and the spread of the ICs. Since differences between ensemble members are indicative of forecast errors, this is a reflection of how quickly errors grow on average. This curve can be thought of as the predictability error growth for a perfect model, given a large  $n$  and properly sampled ICs. The curve does not depend on model error (i.e., model – truth) since it is built from differences between model solutions.

One obvious distinction between Equations (1) and (2) is that the control error variance, Equation (1), is a standard variance calculation over all points, but Equation (2) is an average over all points of the EF variance at each point. This is not however the reason for the gap between the control error and differences curves. The reason is model error. Recall that the ICs are a distribution of possible truths that should contain the true initial state, thus defining many possible initial errors. The differences curve then shows how the variance among these errors increases over the forecast period. The deterministic control forecast begins with one of the initial errors contained in the ensemble, so if a perfect model were used, the control's forecast error (i.e., truth – forecast) would match up exactly with one of the differences within the EF

solutions where both the control member and true evolution of the atmosphere are present. Over a large number of cases, the control's average error variance would naturally match up to the average variance of the ensemble differences. In practice, the control's error variance is larger because model error increases the control's forecast errors. Thus the farther apart the control error and differences curves are, the greater the model error.

The last curve is the variance in the error of the ensemble mean.

$$(s_{\bar{e}}^2)_t = \frac{1}{M} \sum_{m=1}^M ((\bar{e}_{m,t} - o_{m,t}) - \langle \bar{e}^* \rangle_t)^2 \quad (5)$$

where  $\langle \bar{e}^* \rangle_t$  is the average error of the ensemble mean over all  $M$  points at lead time  $t$ . The ensemble mean ( $\bar{e}$ ) for a particular point  $m$  at lead time  $t$  is:

$$\bar{e}_{m,t} = \left( \frac{1}{N} \sum_{n=1}^N e_i \right)_{m,t} \quad (6)$$

The ensemble mean's error variance initially matches the control's, then separates and asymptotes to  $\sigma_c^2$  (Leith, 1974). This is the key to the value of EF. Averaging the ensemble members acts as a very selective filter, smoothing out the nonlinear errors that arose from the erred ICs. The period between the ensemble mean's break from the control forecast to near the  $\sigma_c^2$  asymptote is the period when EF adds value to forecasting of the parameter in question.

The asymptoting behavior of these curves results from the statistics of EF sampling, which can be easily demonstrated. First of all, recall that the forecast PDF evolves to the climate PDF for a very long forecast lead time. The ensemble mean would then exactly equal the climatic mean, and the error variance of the ensemble mean must match  $\sigma_c^2$ . The deterministic control forecast, being a random sample from the forecast PDF, may have an error up to twice as large as the ensemble mean. This of course doubles its variability, making the control's error variance asymptote to  $2\sigma_c^2$ .

To demonstrate these relationships, we simulated  $10^5$  forecasts of *MSLP* by a well-calibrated ensemble of eight members (an EF size of significance to this research) at an extended lead time so that the forecast PDF has spread out to the climate PDF. All samples were drawn from a normal ( $\mu_c = 1011.37$  mb,  $\sigma_c = 10.33$  mb) PDF to simulate the *MSLP* climate PDF in Figure 1. The verification value of *MSLP* used to calculate error for each EF case was a separate random draw from the same PDF. Using Equation (1), we found an error variance of the control to be  $212.82 \text{ mb}^2$  (compared to  $2\sigma_c^2 = 213.52 \text{ mb}^2$ ). Using Equation (2), the average variance of the differences was  $213.84 \text{ mb}^2$ , confirming the  $2\sigma_c^2$  asymptote. Using Equation (5), the error variance of the EF mean was  $119.77 \text{ mb}^2$ , notably higher than the expected  $\sigma_c^2 = 106.76 \text{ mb}^2$ . However, this result must be corrected by a factor of  $n / (n + 1)$  (explained below) to adjust for the small sample size, thus yielding a matched value of  $106.16 \text{ mb}^2$ . This simulation demonstrates that the asymptotic values of the curves in an error variance diagram are a statistical property of sampling from a PDF—the basic process of EF.

The major influence on error growth, which determines the EF value period, is the meteorological scale of the parameter under investigation. For the 500 mb height field (mainly synoptic to planetary variability), Figure 3b shows that EF is useful for medium range forecasting from about day 4 out to at least day 12. It is reasonable to expect that the EF value period for smaller scale, more rapidly varying phenomena of interest to SREF (such as precipitation) should be in the short range. However, this has not yet been clearly demonstrated.

This research involves forecasts to a lead time of only 48 hours, typically well below the limits of predictability. Therefore, we did not make use of the full error variance diagram. Instead, we chose to analyze two other measures of an EF that are by-products of the above quantities. The first metric is the average variance of the EF members about the EF mean as opposed to the average variance of the differences between EF members in Equation (2):

$$\langle s_e^2 \rangle_t = \frac{1}{M} \sum_{m=1}^M \left( \frac{1}{n-1} \sum_{i=1}^n (e_i - \bar{e})^2 \right) \quad (7)$$

This quantity (or its square root, the standard deviation) is referred to in the literature as the *ensemble spread*, and is commonly how the predictability error growth is examined. The second EF metric is the mean square error (*MSE*) of the ensemble mean,

$$(MSE_{\bar{e}})_t = \left( \frac{n}{n+1} \right) \frac{1}{M} \sum_{m=1}^M (\bar{e}_{m,t} - o_{m,t})^2 \quad (8)$$

Note that in an ideal ensemble with large  $n$  applied to a large  $M$ , this quantity matches up perfectly with the error variance of the EF mean (Equation (5)) since the average error of the ensemble mean,  $\langle \bar{e}^* \rangle_t$ , goes to zero.

The importance of these two metrics is that  $MSE_{\bar{e}}$  should match up with  $\langle s_e^2 \rangle$  for a verification that is a random sample from the forecast PDF. This concept, often referred to as *statistical consistency* of an ensemble, was formalized by Talagrand et al. (1999). To put it in plain language, the average difference between the ensemble mean and the verification should be the same as the average difference between the ensemble mean and the ensemble members, so the verification appears to be just like one of the members. Ziehman (2000) pointed out that it is necessary to account for small sample sizes in order for statistical consistency to hold when  $n$  is small. Equation (7) does this by dividing by  $n-1$  in the variance calculation and the standard *MSE* in Equation (8) corrected by a factor of  $n/(n+1)$ . This correction for ensemble size also agrees with Leith (1974).

Comparing  $MSE_{\bar{e}}$  and ensemble spread provides an excellent tool to test for realization of the fundamental goal of EF—production of a forecast PDF from which truth is a random sample. For demonstration purposes, Figure 4 illustrates what we will call a *dispersion diagram*, showing



the common problem in EF of insufficient predictability error growth (i.e., ensemble spread falls short of  $MSE_{\bar{e}}$  and the EF is said to be under dispersive). We will use the term *dispersion* to denote ensemble spread above and beyond the initial spread in the ICs, so it is a measure of how much the members spread out.

An underdispersive ensemble fails to portray the truth and displays overconfidence in its probability forecasts. That is, it over forecasts high probability events and under forecasts low probability events—a clockwise rotated curve on a reliability diagram (see Appendix I). A logical reaction is to increase the IC spread thus producing greater ensemble spread and increased likelihood of portraying truth. However, that is counterproductive for improving the EF if it is done outside the bounds of known uncertainty since that would typically degrade the ensemble's resolution. While we discuss the three basic requirements of EF over the next few sections of this chapter, we will explore factors that lead to low dispersion and possible ways to alleviate the problem and improve skill at the same time.

## 2. Representation of Analysis Uncertainty

The key to successful representation of analysis uncertainty in EF is estimating the true analysis error vector,  $\bar{E}_0$ .

$$\bar{E}_0 = \bar{T} - \bar{A} \quad (9)$$

where  $\bar{T}$  is again the true state of the atmosphere and  $\bar{A}$  is an analysis. It is the fact that we can never know  $\bar{E}_0$  which makes EF necessary. A set of ICs that portrays the true state can be generated by making perturbations about a best guess analysis based on an estimate,  $\hat{\bar{E}}_0$ .

An estimate of the magnitude of the analysis error may be obtained by comparing many analyses and re-analyses (made at a later time with additional observations). Perturbations about

the best guess analysis may then be made in random directions (Figure 5a) scaled by that magnitude, creating a cloud of ICs symmetric in all dimensions. This so-called Monte Carlo (Leith, 1974) approach is theoretically effective for a very large number of ICs but is extremely inefficient since the majority of the members yield repeated information. Since most of the perturbed ICs lie off the attractor, during the forecast evolution their trajectories converge toward the members' trajectories that started on the attractor. While this method should generate realistic predictability error growth allowing the forecast PDF to portray the true state, it is not practical for operational EF because of the large processing cost.

Having an estimate, or estimates, of the analysis error vector allows for an efficient EF (Figure 5b). Such an estimate is actually a two-way vector with the best guess analysis at the center, again scaled by the time average analysis error in each direction. Since the true state should lie close to this vector, ICs are placed along it. All these ICs are on or very close to the attractor so their trajectories diverge and yield very different solutions which should portray the future true state. One risk of this method is that putting too much confidence in the  $\hat{\mathbf{E}}_0$  can throw the system off if the estimate is poor. Another risk is the possibility of oversampling part the PDF if the ICs are placed too close together.

Most of the EF research effort over the past decade concentrated on producing a good  $\hat{\mathbf{E}}_0$  for MREF (Stensrud et al., 2000). This has resulted in successful medium-range EF systems such as the NCEP MREF (Toth and Kalnay 1993; Tracton and Kalnay, 1993) and the ECMWF Ensemble Prediction System (Molteni et al., 1996). At NCEP the estimates are made by the method of breeding of growing modes while ECMWF uses singular vectors. Both of these methods are designed to generate maximum dispersion among the ensemble members several days into the forecast period. This may be a good idea for representing the tails of the forecast PDF but may not produce the complete PDF.

The focus on applying well-formulated ICs in the early MREF systems was likely the result of two factors. First, Lorenz (1963, 1969) clearly showed that a dynamical system is sensitivity to ICs. Secondly, Downton and Bell (1988) explained that for MREF in the midlatitudes the perfect model assumption is reasonable. So it was logical and productive for an EF such as the NCEP MREF to ignore the complications of model uncertainty.

#### a) Methods for Analysis Uncertainty Representation

In this section we will briefly review the six primary methods (three of which are used operationally) for estimating the analysis error vector and generating a set of ICs for an EF system. There is an extensive body of literature comparing and contrasting the merits of the various methods for MREF, for which these methods were primarily designed. There has been no published research on which method may work best for SREF, whether any are appropriate at all, or whether some new method is required. We will simply present the basic methods to justify and put our choice (multianalysis) in perspective.

As discussed above, the pure Monte Carlo method of generating many random perturbations is impractical because it requires an extreme amount of computer power to get good results (Lorenz, 1993; Wilks, 1995). A modified Monte Carlo method was developed by Errico and Baumhefner (1987). Rather than having many members with totally random perturbations (equal noise at all scales), they thought it would be more efficient and effective to have scale-selective perturbations. Perturbations can still be random, but the amount of noise added to the control analysis at each wavelength is based upon the suspected uncertainty at that scale. The scale-selective perturbations are created by manipulation of the spectral decomposition of random perturbations.

One important finding of Errico and Baumhefner (1987) with consequences for our research is that "...the forcing of small scales by large scales is substantial." They found that, when only

the small scales were perturbed, the ensemble solutions were very similar at all scales, but when they perturbed only the large scales, the ensemble solutions varied at all scales. A second important finding for our research is that, when using a limited-area model (LAM), it is important to perturb the lateral boundaries as well as the ICs. Ignoring that fact leads to limited predictability error growth.

The method of Errico and Baumhefner (1987) has only been used as a research tool, as in a SREF study by Du et al. (1997) in which a 25-member ensemble was run for a single forecast case of explosive cyclogenesis using MM4 at 80-km resolution and ICs as described above. The focus was on producing a quantitative precipitation forecast (QPF). They found that short-range QPF was very sensitive to analysis uncertainty for explosive cyclogenesis, indicating rapid error growth. They also reported that 90% of the root-mean-square error (*RMSE*) improvement by the EF mean was found with 8-10 members, thus confirming the correction factor of Equation (8). Finally, they made the tentative conclusion that "SREF can now provide useful QPF guidance and increase the accuracy of QPF when used with current analysis-forecast systems." This was encouraging for SREF but hardly convincing considering the limitations of the study.

A second method for IC generation, also with a Monte Carlo element, is commonly called Perturbed Observations (PO) and was developed by Houtekamer and Derome (1995) for operational use in the Canadian Meteorological Centre's (CMC) EF system. This method assumes that most of the error in an analysis comes from the errors and incomplete coverage of the observations. To generate another likely analysis, random errors (consistent with known error characteristics) are added to the observations followed by another separate analysis cycle. This can be repeated  $n$  times to produce  $n$  analyses. Currently, the CMC's ensemble consists of 8 members run with a global spectral model and 8 members run with their Global Environmental Multiscale (GEM) model.

A significant limitation to the PO approach is that the differences between the analyses are limited by the fact that they all use the same processing methods in their analysis cycle. Errors introduced by using the same model for the first guess as well as the same optimal interpolation scheme make all the ICs share similar deficiencies. A more complete PO method would account for the uncertainty in both model and optimal interpolation scheme within the analysis cycle. Nevertheless, Hamill et al. (2000b) showed that the PO method is superior to the other two operational methods discussed next.

A third method, called Singular Vectors (SV), generates ensemble ICs mathematically (Molteni et al., 1996). It uses the idea of Lorenz (1965) who proposed that "optimal perturbations" that grow the fastest in the short-range are revealed by the largest eigenvalues of the eigenvectors (i.e., singular vectors) of a symmetric error covariance matrix (i.e., a description of the forecast error PDF). The justification for using these modes as perturbations to the best guess analysis is that since only a limited ensemble can be run, choosing the fastest growing modes should ensure that the true evolution of the atmosphere is consistently portrayed. The SV method went operational in the ECMWF Ensemble Prediction System (EPS) in 1992, and is presently run with 51 members at T255L40. To find the symmetric covariance matrix, an adjoint (i.e., linear tangent) version of the global model is used to find the maximum growth at 2 days.

There are several notable problems with the SV method. One is that it is computationally expensive to find the optimal perturbations. Secondly, since it can only examine linear error growth, it is limited to maximizing the 2-day growth. It is unreasonable to expect the optimal perturbations at 2 days to continue to be the fastest growing modes into the medium range. Lastly, SV by design tends to sample the extremes of the analysis PDF instead of providing a purely random sampling. The sampling is also limited by the fact that not all the errors present in the analysis project onto growing modes.

A fourth method called Breeding of Growing Modes (BGM), described by Toth and Kalnay (1993), was developed for the NCEP global ensemble and is actually a clever, efficient approximation to SV. The basis of this method is that while an analysis cycle is designed to produce acceptably small errors, the largest differences between the analysis and truth are believed to project onto growing modes because of the use of the model first guess in objective analysis. The BGM method produces alternative ICs by mimicking an analysis cycle. To make  $2n$  ICs, BGM begins by making  $n$  unique, random perturbations (taken as + and -) to the best guess analysis. A short forecast from each perturbation is compared to the next best guess analysis to provide an estimate of a growing mode, which is then scaled to provide a perturbation for the ensemble. The NCEP global ensemble currently consists of 24 members run with the Global Forecast System (GFS) model at T126L28 (T62 after 84 h). The BGM method suffers from the same basic problem as SV in that it tends to reflect the extremes of the analysis PDF. Additionally, Baumhefner (2000) demonstrated that the members are highly correlated and that their differences do not resemble the typical analysis error structures.

The fifth, and perhaps most promising IC method for SREF, is the Ensemble Kalman Filter (EnKF). It has yet to be applied operationally, but is described by Hamill and Snyder (2000c) primarily as a means to improve the analysis. One of the weakest parts of any analysis scheme is poor knowledge of the error in the first guess field (represented by an error covariance matrix) that is to be combined with observations. A true Kalman filter would find the error covariance matrix directly through linear dynamics, and is, for practical purposes, computationally impossible for the degrees of freedom in NWP. The EnKF method assumes that an approximation to the matrix can be provided by an ensemble of short-range forecasts, run parallel to the analysis cycle, that applies the PO method. This process produces a greatly improved analysis by minimizing the error based on the sensitivity to the variance in the observations.

More importantly, it also produces a set of ICs specifically conditioned for SREF since a large component of short-range forecast errors come directly from the error in the first guess field, which EnKF captures. One limitation to EnKF is that a large ensemble is required to properly represent the error covariance matrix (Hamill and Snyder, 2000c). It is also unclear how model error may impact the EnKF process.

The sixth IC method, which is applied in our research, is termed multianalysis and was developed by Grit and Mass (2002) as a research tool. It can be considered semi-operational since it has been run in real time since its inception in January 2000. The multianalysis method uses several independent, large-scale analyses/forecasts produced from different forecast centers to initialize and provide lateral boundary conditions for a mesoscale model. In essence then, the goal of this process is to take the original synoptically diverse solutions and project them down to the mesoscale, thus producing a SREF with differences that should estimate likely mesoscale errors.

One key assumption then is that the differences among the analyses are representative of analysis error. The other assumption, supported by Errico and Baumhefner (1987), is that the largest component of the mesoscale forecast error actually originates from synoptic-scale errors in the analysis. The validity of these assumptions, as well as all the limitations for the multianalysis method, will be discussed in Chapter II. The vital fact is that Grit and Mass (2002) showed that running a multianalysis ensemble with only five MM5 members provides excellent prediction of forecast error on the mesoscale. It was this result that led us to use this method to define the ICs for our research. All of the other methods for representing analysis error have questionable applicability to SREF or are beyond our computational capabilities.

### 3. Representation of Model Uncertainty

There is evidence that even in the medium range, discounting model uncertainties leads to inferior EF performance (Buizza et al., 1999; Harrison et al., 1999). This likely applies even more for SREF where the impact of model error can be amplified (Brooks and Doswell, 1993). In short-range, high-resolution forecasting of sensible weather phenomena, the model is highly sensitive to its parameterizations (Stensrud et al., 2000). Therefore, regardless of IC quality or ensemble size, a SREF system that ignores model uncertainty can not generate the proper predictability error growth. While there is still much research to be done on accounting for model uncertainty (Hamill et al., 2000), its inclusion appears necessary for construction of an effective SREF.

Use of parameterizations within an NWP model play a large role in limiting ensemble dispersion when their errors are not accounted for. These parameterizations are best estimates of quantities, factors, or processes that are either not completely known or of too small a scale to be resolved by the model. In nature, the phenomenon (estimated with a constant in the model) is often highly variable over space, time, and different weather regimes. A parameterization may reasonably represent some natural process at times and poorly represent it at other times. When the members of an ensemble all use the same limiting parameterizations, they all evolve with the same model errors thus failing to account for model error. The members' similarities result in an underdispersive system.

Limited ensemble dispersion, commonly seen in ensemble systems (Buizza, 1997; Hamill and Colucci, 1997; Tallagrand, 1999), may be due to either ICs which insufficiently project onto growing modes or to poor representation of model uncertainty (Buizza, 1995; Houtekamer, 1996). We should imagine then that the total ensemble dispersion can be defined as the increase in ensemble spread from its initial value and that both IC error and model error are



simultaneously contributing to the dispersion (Figure 4). Our hypothesis is that in the short range, the percent contribution to ensemble dispersion from model uncertainty may be as big, or bigger than from IC uncertainty, depending upon the parameter and scale. This does not contradict the fact that dispersion in the medium range results mainly from analysis error, which is another way of saying that predictability limits are primarily determined by analysis error (Lorenz, 1963; Ziehmann, 2000).

The set of ICs is critically important for defining the initial envelope of solutions, but the differences between the members take time to grow. Generally, their growth is relatively weak and linear in the first 24 h and is followed by increasing nonlinear growth once they become well organized (Gilmour, 2000). Model errors typically have high spatial variability so have very little large-scale structure to project onto growing modes. Also, model errors do not have to organize before growing, so they reach their peak influence on the solution shortly into the forecast cycle. Inclusion of model uncertainty in a SREF should therefore significantly improve ensemble dispersion, creating a much better estimate of predictability error growth in the short range. Furthermore, model parameterizations have the greatest impact on the solution at or near the surface (Mullen and Baumhefner, 1988), so the best way to improve SREF is to introduce model perturbations (i.e., variations to model parameterizations) that focus on increased variance in surface and sensible weather variables (Stensrud et al., 2000).

Another hypothesis of this research, supported by Mylne (2002), is that the addition of model perturbations to a SREF can increase dispersion and simultaneously improve the resolution component of *FP* skill. Addition of model perturbations does not simply add arbitrary spread to an underdispersive ensemble, but rather it introduces actual uncertainty to the ensemble that was previously omitted. The correct but difficult solution to the low dispersion problem of EF is to thoroughly represent both IC and model uncertainty.

It has been proposed that model deficiencies can be separated into two distinct classes, namely *systematic* and *stochastic* (Hamill et al., 2000a). Systematic error refers to model bias and is normally blamed on poorly tuned parameterizations. However, Buizza (1999) describes how parameterized physical processes can lead to random (i.e., stochastic) error without any bias. This is possible because over many model time steps a parameter may accurately represent the average value of some sub-grid scale physical process (e.g., precipitation droplet growth) but be in error by a random value at any one time step. With that in mind, we will use the term *systematic error* to simply mean model bias (i.e., any forecast error that regularly reoccurs) and not tie it to any particular source. *Stochastic error* is then the remainder of the forecast error (which is random) and is also what we refer to with the term *model uncertainty*.

The sources of model error can be broken up into three basic categories (Table 1), where each category requires different basic methodology for inclusion in a SREF. Each source among the categories may contribute differently to both systematic and stochastic error. The term *physics parameterization* will be used in reference to a model's estimation of a poorly known and/or unresolved quantity or physical process. The error in a physics parameterization could be represented by perturbing about its estimated uncertainty during model integration. A *surface boundary parameter* (SBP) is like a physics parameterization in that for a single model run, it is an estimate of the average value of some poorly resolved quantity. The difference is that it is also spatially dependent so should therefore be perturbed about its estimated uncertainty over the entire model domain. Lastly, the *numerical processing* model error category contains the errors associated with the mathematics of NWP and its application to computers. Perturbing about these errors is not straightforward since it would be very difficult to represent such error and maintain equality among ensemble members.

Our research compares and contrasts two different strategies for representing model uncertainty in a SREF. Both use the multianalysis method of defining the set of ICs. One strategy, commonly termed the *multimodel* approach, is to use more than one NWP model for the ensemble members. Each member then has a unique IC and may have a unique model or share a model with some other members (depending on the number of models applied). Our application of this approach is called a *multimodel multianalysis* (MMMA) ensemble. The other strategy, called the *perturbed-model* approach, is to apply a set of ICs to just a single model framework but use many different versions of, or perturbations to, that model. Each ensemble member then has a unique IC and a unique, but related, model. Our application of this approach is termed a *perturbed-model multianalysis* (PMMA) ensemble.

A hypothesis of this research is that while the PMMA approach may be more theoretically proper since it attempts to more rigorously account for model error, the MMMA approach is more practical to employ and produces as good or even better results. When using an imperfect model in an EF, it is imperative for the members to have various model attractors that bound the true attractor; otherwise, the members will not be drawn from the forecast PDF (Hansen, 2002). In the MMMA approach, each member has a model with a drastically different model attractor that provides unique skillful information to the ensemble (Evans, 2000). The spread among the various model attractors may be a reasonable representation of model uncertainty. In PMMA, each member has a unique model but many of the same model aspects are shared. The resulting set of model attractors may be too constrained to fully represent the uncertainty about the true atmospheric attractor.

#### a) Perturbed-Model Theory

Since one of the main efforts in this research was the construction and implementation of a PMMA, we need to discuss the theory of model perturbations in more detail. The challenge of

the PMMA strategy of representing model uncertainty is that the sources of error within the model are numerous, mostly unknown, and vary greatly in character. Attempting to completely and accurately represent all these errors individually in a SREF would be a daunting task.

In the ideal PMMA all aspects of model uncertainty are rigorously represented. This could be done by defining the uncertainties with PDFs for all the parameterizations and numerical imprecision. An ensemble of distinct and equally likely models could then be made with various combinations of random samples from all those PDFs. (Note that we can think of deterministic-style forecasting as using only the expected value of those PDFs.)

Defining parameterization PDFs would certainly be a challenge, but the real difficulty comes when trying to capture all the model uncertainty. To thoroughly span the space of models, it would be necessary to run all possible combinations of the various parameter values from the PDFs. Even if only a few samples are taken from each PDF, the limits of today's computer systems are quickly exceeded. We can compute the number of required model runs ( $M$ ) by

$$M = N \prod_{i=1}^A p_i \quad (10)$$

$$M = N p^A \quad \dots \text{for constant } p$$

where  $N$  is the number of ICs,  $A$  is the number of distinct aspects of the model being perturbed, and  $p_i$  is the number of unique perturbations per model aspect (like samples from a distribution).

Figure 6 gives a simple example where we start with only 4 ICs and take just 2 random samples each of 3 different physics parameterizations. To capture all the possible combinations, which may all be equally likely, we need to process 32 ensemble members. This is similar to a decision tree in statistics: each branch of the tree is an ensemble member with a different and equally likely version of the model. However, the various members are only all equally likely if the model perturbations are independent and uncorrelated.

The tree diagram for a more thorough system would be gigantic. Say we use a set of 10 ICs and we identify 20 distinct aspects of the model that may be in error. We also choose to generate 30 perturbed values (i.e., samples) for each of those model aspects—the standard minimum number of samples to represent a PDF. Equation (10) gives a total of about  $3.5 \times 10^{30}$  required model runs! The only practical option is to use a very limited subset of those runs, as in the work by Houtekamer et al. (1996) discussed in the next section. The point of the PMMA strategy is not to thoroughly represent model uncertainty since that may be impossible, but rather to run each member with a perturbed model in order to realistically increase ensemble dispersion. Furthermore, for practical constraints only one tree branch is used for each IC but the branches are as different as possible. For example, a four-member PMMA ensemble from Figure 6 might be A111, B122, C212, and D221. While this is an extreme approximation to the ideal PMMA, its efficiency may actually make sense. Running the complete set of model variations would likely waste processing time since similar perturbation combinations would yield very similar solutions. (E.g., model runs A111 and A112 would likely be nearly identical). Selecting only one model variation per IC should efficiently provide additional dispersion and allow the SREF to represent a significant portion of model uncertainty.

#### b) Research into EF and Model Uncertainty

Epstein's (1969) formulation of a stochastic dynamic forecast model was designed to incorporate model uncertainty into the prognostic equations but is impractical for NWP since the equations are unmanageable by numerical methods. Leith (1974) proposed the idea of ensemble forecasting as an approximation to stochastic dynamic forecasting, focusing primarily on IC error or what he termed "internal error." However, he did point out that

"...there is an additional external error generated by the discrepancy between the dynamics of the model and that of the real atmosphere arising in part from the limited dimensionality of the model phase space."

We can only surmise that Leith (1974) chose to assume a perfect model for his study because he believed analysis uncertainty dominates error growth, or perhaps the idea of including model uncertainty was too overwhelming.

The pioneering effort for representing model uncertainty in an EF was Houtekamer et al. (1996), who employed a limited perturbed-model approach. Using a spectral model (T63/L23), the perturbed observation approach was used to create a set of ICs for a low-resolution, medium-range ensemble system with forecasts out to 15 days. A so-called system simulation experiment (SSE) method was then applied to represent model uncertainty using many different versions of the same model. (Note that this EF is different from our PMMA in that the set of ICs was generated "in house" using one analysis system. The PMMA employs various analyses from different forecast centers' analysis systems.)

Houtekamer et al. (1996) ran two different ensembles of eight members each. One ensemble used a unique IC but the same model version for each member, while the other used a unique model version for each member as well as a unique IC. The setup for each of the eight model versions was chosen from four model options (horizontal diffusion, convection/radiation, gravity wave drag, and orography) with two choices each and three SBP options (sea surface temperature, roughness length, and albedo), each with eight different choices. Applying Equation (10), the eight members then represent only eight possible perturbed model combinations out of  $8(4^2)(3^8) = 839,808$ . But even with this limited sampling of the model uncertainty, there was notably increased dispersion.

The goal was to correctly boost the predictability error growth of an underdispersive medium range EF that previously had no model perturbations. Figure 7 (using data from Houtekamer et al., 1996) shows that, while the dispersion was increased, the effect was quite limited in the medium range. Houtekamer concluded that while including model perturbations does improve an

EF, "more dramatic perturbations to the model" would be required to produce better error growth. (The similarity of Figure 7 and Figure 3 is not coincidental. Recall that EF spread behaves just like the *MSE* of the EF mean in an error variance diagram, a point that we exploit in our analysis.)

It is difficult to determine the implications for SREF from these results since the model and resolution were geared toward the medium range. In addition, the study only used a single forecast case so the generality of the results is unclear. One curious fact that was revealed (and not discussed since they were primarily concerned with the medium range) is an indication in Figure 7 that model uncertainty may play a much bigger role in the short range by contributing a large part of the dispersion. (Recall that our definition of dispersion is EF spread above the initial spread). Beyond the short range, nonlinear error growth from synoptic-scale differences in the ICs dominates the dispersion, and model errors only add slightly more spread. In the short range, model perturbations quickly make significant differences in the solutions.

Stensrud et al. (2000) performed a study using the perturbed-model approach that did discuss the relationship between forecast lead time and error growth by model error. They compared the behavior of two very different, 19-member SREF systems using the MM5. The "IC ensemble" had perturbed ICs defined with the Mullen and Baumhefner (1988) approach, all using the same model. The "physics ensemble" used one IC but 19 different versions of MM5 defined by 5 convective scheme options, 2 boundary layer options, and 3 levels of moisture availability. This was an interesting way to isolate the error growth due to model error, but it is not appropriate to do a skill comparison of these two ensembles since the physics ensemble is unfairly degraded by a lack of representation of analysis error. Another serious limitation is that this study only examined one complete forecast case.

Nevertheless, this work of Stensrud et al. (2000) does provide some evidence for the key idea that use of perturbed models gives the ensemble members different systematic errors, thus

providing a more appropriately diffuse forecast PDF. They showed that the spread in the physics ensemble grew two to six times faster in the first 12 h compared to the IC ensemble. They also concluded that the influence of model uncertainty on forecast error is largest in the short-range for meteorological variables at the surface. Stensrud et al. (2000) summarized the benefit of the perturbed model technique:

“By using different models, in conjunction with different initial conditions, it may be possible to increase the accuracy and usefulness of an ensemble by creating greater divergence in the ensemble trajectories than would be created by using only different initial conditions.”

The studies of both Houtekamer et al. (1996) and Stensrud et al. (2000) were a major influence on the choice and design of the PMMA technique applied in our research.

An alternative to either the multimodel or perturbed-model approach, called stochastic physics, was applied by Buizza et al. (1999). The basic assumption with Buizza's method is that random errors coming from the various parameterizations have a high degree of spatial and temporal coherence and that the errors are proportional to the tendency (i.e., rate of change). Instead of trying to handle all the errors separately, stochastic physics attempts to capture their influence by randomly perturbing the tendency of state variables with some appropriate degree of spatio-temporal autocorrelation.

Buizza et al. (1999) did find that this method increased ensemble spread and improved performance, but others have found that stochastic physics fails to represent the full spectrum of model uncertainty (Evans et al., 2000; Ziehmann, 2000; Richardson, 2001a). Forecasters have found that subjectively, the differences among the ECMWF EPS members (that use stochastic physics) fall well short of the synoptic differences found from a multimodel ensemble (Mylne et al., 2002). The limits of stochastic physics may be due to the fact that all the members use the same model attractor. The random perturbations give occasional kicks off the attractor to the



trajectories, which then quickly reconverge. The effect then is that the solutions are still very similar. For all these reasons, we chose not to apply stochastic physics in our SREF research.

A SREF study that explored several questions of relevance to our research was accomplished by Wandishin et al. (2001). They compared error growth and skill over a relatively large sample of 43 total cases (27 cool season and 16 warm) of several subset ensembles of the NCEP SREF—a 15-member MMMA ensemble (Du and Tracton, 2001). A weakness of this study is that the NCEP SREF uses ICs that likely do not adequately represent analysis errors. The 15 members consist of five Eta model runs that use multianalysis ICs (all produced at NCEP and thus highly correlated), five more Eta runs that use bred-mode ICs, and five Regional Spectral Model (RSM) runs using the same bred-mode ICs. Use of two models does provide an element of multimodel representation of model error, but the system is seriously encumbered by the poor ICs.

Wandishin et al. (2001) conceded that their study was rather limited and that “future work is needed to quantify the roles of model formulation and initial condition uncertainty.” They did conclude however that SREF can give useful guidance on probabilistic QPF whereas information from a deterministic forecast is quite limited and much less useful. Additionally, relevant to our research, they found that error growth for mesoscale parameters is very weak compared to the growth found in a synoptic-scale parameter such as 500 mb GPH. They did not address whether the weak error growth was due to error saturation, verification method, or some other effect.

A study by Evans et al. (2000) provides some insights into the value of the multimodel technique for representing model uncertainty. Focusing on MREF with 9 cases, they compared the skill of three, 34-member ensembles: 1) a random subset from ECMWF’s 51-member EPS that used stochastic physics, 2) an ensemble that used the same ICs as the ECMWF EPS but used the United Kingdom Meteorological Office (UKMO) global model, and 3) a combination of 1 & 2, using 17 members from each.

They came to the dramatic conclusion that the multimodel ensemble outperformed the ECMWF EPS in both deterministic skill of the EF mean and in the skill of probability forecasts. This improvement was not simply due to adding more members or forcing spread toward the climate PDF, since they found improvement in both reliability and resolution. It was likely "due to the sampling of different, skillful populations provided by the individual systems." They also concluded that for the medium range, model errors do contribute significantly to the total forecast error so must be accounted for in an ensemble system.

The benefits of a multimodel ensemble in the medium-range were further demonstrated by Ziehmann (2000). Over a large sample of forecasts (90 cool season and 90 warm season), she compared a random subset of four ECMWF EPS (with stochastic physics) members to a 4-member poor man's ensemble (PME). The conclusion was that not only does the PME beat the ECMWF EPS subset but that it even beat the full 51-member ECMWF EPS in several key aspects of EF performance.

Ebert (2001) explored the PME to see how a seemingly nonrigorous EF method can be so effective. Using a 7-member ensemble comprised of global models from various operational centers to examine the skill of QPF, she also found ensemble superior to the 51-member ECMWF EPS. Ebert noted that:

"Because it [PME] samples uncertainties in both the initial conditions and model formulation through the variation of input data, analysis, and forecast methodologies of its component members, it is less prone to systematic biases and errors that cause underdispersive behavior in single-model ensemble prediction systems."

She also concluded that for probabilistic forecasts, there was no need for a calibration such as applied by Eckel and Walters (1998). However, the dispersion of the PME was not investigated (i.e., improper dispersion indicates a need for calibration) so her conclusion is really a hypothesis, which was explored in our research (see Chapter III).

An expanded version of the study by Evans et al. (2000) was conducted by Richardson (2001a) to examine the possibility of improving upon the ECMWF EPS through inclusion of multianalysis and/or multimodel techniques. This study included 60 forecast cases (mostly cool season) and compared five different ensembles: 1) the 51-member ECMWF EPS, 2) the 27-member UKMO ensemble that used the same ICs as the ECMWF EPS, 3) a 54-member combination of 1 & 2, using 27 members from each, 4) a 55-member multianalysis ensemble made by applying 11 ECMWF EPS perturbations each to analyses from 5 different centers, and 5) a 51-member ensemble made by applying the ECMWF EPS perturbations to a "consensus analysis" (This is what we will call the *centroid* analysis).

Richardson also found that the multimodel ensemble beat the ECMWF EPS but added that a comparable improvement was realized by the multianalysis ensemble. What helped him reach this conclusion was the removal of bias from the model output. Model bias can be a significant part of the forecast error and should be removed before considering the ensemble of forecasts, but curiously it is regularly ignored in most EF studies and applications. Richardson showed that bias removal improved the skill of the EF mean and the probability scores of his EFs and also allowed for a more equitable comparison of ensemble systems that employ different models. This key idea was adopted and explored in our research.

The most extensive study to date concerning the benefits of a MMMA in the medium range was accomplished by Mylne et al. (2002). Using 75 cool season and 85 warm season cases, they followed the basic method of Evans et al. (2000), but their MMMA consisted of 54 members (27 members from the ECMWF EPS and 27 UKMO model runs), which could be directly compared to the full ECMWF EPS. Their conclusion was that the MMMA improved upon the skill of the ECMWF EPS by about 10%. (This is yet another example of the deficient representation of model error by stochastic physics in the ECMWF EPS.)

Mylne reasoned that "...the benefits of whichever is the better system at a particular time and place may be obtained all the time through better probabilities." In other words, one model may be superior overall but the relative skill among the models shifts over time and space. Only a member that is consistently inferior can add no value to the ensemble system. The notion of using unequally skilled members appears to go against the conventional wisdom of EF that members have to be equally likely to be considered random samples from the forecast PDF. In our research, we sought to resolve this matter.

To conclude, this literature review presented the source of many of the ideas and methods that we applied in our research. One important issue not addressed in the EF literature to date is the theoretical differences and relative merit of the perturbed-model vs. the multimodel approach for representing model uncertainty. This is another major question we addressed in our research.

#### **4. Sufficient Ensemble Size**

The requirement of sufficient ensemble size is much more straightforward compared to accounting for analysis and model error, but it is no less critical. Since more members makes for a better EF, one would like to run an EF system with many, many members. Unfortunately, current computer capabilities limit the size of an operational EF to well below what is required for thorough sampling. In our research, the number of members was also constrained by choosing to use independent analyses as ensemble ICs, of which there is a limited supply.

It is very important to understand the impact that ensemble size has on EF performance for two reasons: 1) the ensemble size must be considered in designing a system to be of value to specific applications, and 2) to properly analyze the skill of a particular EF methodology, the deficiencies caused by low sampling should not cloud the analysis.

Generally, EF performance decreases as ensemble size decreases (Buizza and Palmer, 1998; Richardson, 2001b), but the impact of this effect depends upon what aspect of an ensemble is

considered. The skill of the EF mean is only minimally affected, a fact highlighted by Du et al. (1997), who confirmed that 90% of the benefit of the EF mean can be achieved by an 8-10 member ensemble. This is a good example of separating out the deficiencies caused by low sampling. Recall that in the discussion on predictability error growth we had to make a correction to the *MSE* of the EF mean by a factor of  $n/(n+1)$  in order to arrive at the theoretical *MSE* value for  $n = \infty$ . The smaller the sample, the lower the skill of the EF mean. The reason for this is explained further below.

*FP* is severely affected by low sample size. In fact, Richardson (2001b) went so far as to say "An ensemble of ten or so members should not be expected to provide reliable probability forecasts." We strongly disagree with this statement on the basis that studies using ensembles with 10 or fewer members have demonstrated skilled *FP* (Ziehmann, 2000; Ebert 2001). Additionally, Richardson (2001b) made an error in his research (discussed in Chapter II when we cover how to calculate *FP* from an EF). Nevertheless, we do agree with Richardson's conclusion that a large ensemble is required for highly skilled *FP* and that the result of low sampling is an overconfident EF. (I.e., the PDF tails are less likely to be represented, so high *FP* values are normally overforecast and low *FP* values are normally underforecast.)

The basic problem with a small ensemble is that it can not produce a consistent PDF, often misrepresenting the distribution from which it was drawn. Say we have an ideal ensemble that could produce a perfect analysis and forecast PDF when sampled infinitely so that the true state is always a random sample of the ensemble's PDFs. Using only a finite number of members from this same ensemble, we get an approximation to those perfect PDFs, thus harming our EF. The approximate PDFs can still turn out to be excellent, but the smaller the ensemble the more infrequent this becomes and the more unreasonable the approximation can get. This is a basic property of statistics which can not be avoided in ensemble forecasting.

Since this problem arises from statistics, the theoretical implications can best be explored by examining the how the sampling distributions of a PDF's moments (sample mean  $\bar{x}$  and sample variance  $s^2$ ) change with increasing sample size. A sampling distribution is produced by generating  $M$  ensemble realizations with a fixed ensemble size  $n$ , then plotting the  $M$  values of  $\bar{x}$  and  $s^2$ . Apart from using a finite  $n$ , we assume an ideal ensemble so all EF members are drawn from the correct forecast PDF.

There are two relevant questions. How much error in the moments can we expect from any one ensemble realization? On a long-term average basis (i.e., after many forecast cases so there is a large sample) do  $\bar{x}$  and  $s^2$  produce good estimates of their theoretical values? For each simulated ensemble of  $n$  members we repeatedly generated  $n$  random samples from a forecast PDF (defined below) and compared the sample statistics (which will naturally have some error) to the population mean  $\mu$  and population standard deviation  $\sigma$  of the forecast PDF. This sampling experiment mimicked an ideal ensemble's effort at representing the forecast PDF. While such an examination may seem oversimplified, it is actually very applicable to our complex problem of sampling the high degrees of freedom of the atmospheric PDF. We can think of a forecast PDF as a multidimensional collection of many single-variable PDFs, one for each state variable at every grid point. So the basic arguments presented here for a single-variable PDF should extend to the entire forecast PDF.

The sampling distribution of  $\bar{x}$  follows a normal distribution, regardless of the governing PDF. In the long-term, the expected value of  $\bar{x}$  converges to  $\mu$  according to the Central Limit Theorem (Devore, 1995).

$$E(\bar{x}) = \mu \tag{11}$$

This is a very good thing for ensemble forecasting since it means that, over many forecast cases, the EF mean matches the PDF mean no matter what  $n$  is. The amount of possible error in any one

ensemble realization is determined by the variance of  $\bar{x}$ , which is equal to the forecast PDF's variance divided by the ensemble size.

$$V(\bar{x}) = \frac{\sigma^2}{n} \quad (12)$$

or, equivalently, the standard error is  $\sigma/\sqrt{n}$ . So as stated previously, the error in the EF mean decreases with increasing  $n$ .

The sampling distribution of  $s^2$  follows a  $\chi^2$  distribution. In the long-term average, the expected value of the sample variance converges to the true variance.

$$E(s^2) = \sigma^2 \quad (13)$$

This is another very desirable fact for EF since it means that, regardless of  $n$ , the EF spread matches the forecast PDF's variance in the big picture. The variance of  $s^2$  is not as straightforward, involving many higher order moments. For our purposes here, we simply calculated the variance of  $s^2$  empirically for our assumed PDF, thus determining the amount of possible error in  $s^2$  for any one ensemble realization.

The governing forecast PDF was defined as the standard normal,  $N(0,1)$ , to make the results generic and normalized about  $\sigma = \sigma^2 = 1$ . The top graph in Figure 8 shows sampling distributions of  $\bar{x}$  for  $n = 8$  and successive doublings ( $n = 8, 16, \dots, 1024$ ). The distributions are all centered about the population mean ( $\mu = 0$ ) as in Equation (11), but we are more interested in the possible error for any one case. Larger errors are of course more likely for distributions with greater variance, corresponding to the smaller sample sizes as described by Equation (12).

Dividing the standard error by  $\sigma$  gives the normalized standard error ( $\bar{x}_{NSE}$ ) of  $1/\sqrt{n}$ , plotted as the solid curve in Figure 9. The typical value of the erred mean ( $\bar{x}_{typical}$ ) for any PDF is then found by:

$$\bar{x}_{typical} = \mu_{true} \pm \bar{x}_{SE} \quad (14)$$

where  $\bar{x}_{SE} = (\bar{x}_{NSE})\sigma_{true}$  is the magnitude of the standard error in the mean and  $\sigma_{true}$  is the forecast PDF's standard deviation.

The bottom graph in Figure 8 shows sampling distributions of  $s^2$ . The distributions are all centered about the true variance ( $\sigma^2 = 1$ ) as in Equation (13). Notice that for large values of  $n$ , the  $\chi^2$  distribution approaches a Gaussian distribution but for low  $n$  there is a wide, heavily skewed distribution. Points on the dashed curve in Figure 9 were found empirically using 5000 sample draws from  $N(0,1)$  with a fixed  $n$  to produce a  $\chi^2$  distribution from which variance of  $s^2$  was then calculated. The results are automatically normalized to  $\sigma^2 = 1$  so the typical value of the erred variance ( $s_{typical}^2$ ) for any normal PDF is found by:

$$s_{typical}^2 = \sigma_{true}^2 \pm s_{SE}^2 \quad (15)$$

where  $s_{SE}^2 = (s_{NSE}^2)\sigma_{true}^2$  is the magnitude of the standard error in the variance,  $s_{NSE}^2$  is the normalized standard error in the variance and  $\sigma_{true}^2$  is the forecast PDF's variance.

The implications of undersampling to ensemble forecast are now clear. For small  $n$ , the typical error in the mean is a significant portion of the forecast PDF's standard deviation, causing a notable shift in the estimated PDF. The typical error in the variance for small  $n$  is an even larger portion of the correct variance, causing a prominent squeezing or stretching in the estimated PDF. A larger ensemble has narrower sampling distributions and an improved ability to consistently reproduce the PDF from which the members are drawn, thus improving ensemble skill. It appears that the most significant improvements should be expected as the number of members is increased into the 50 to 100 range since the standard errors decay exponentially. Beyond that, improvement becomes minimal as more members are added.



So even with a well designed EF system, undersampling alone can result in a poor estimation of predictability error growth. It is therefore quite encouraging that the rather small, experimental SREF systems to date displayed some value and skill (Du et al., 1997; Stensrud et al., 1999; Wandishin et al., 2001; Grit and Mass, 2002). By expanding upon these prototype SREF systems with improvements in ICs, model error representation, and larger ensemble size, we believe that a valuable SREF system is possible. In the next chapter, we will discuss our design and implementation of such a system. We will also revisit this simplified sampling experiment in more depth to examine the likely impact of ensemble size to our SREF systems.

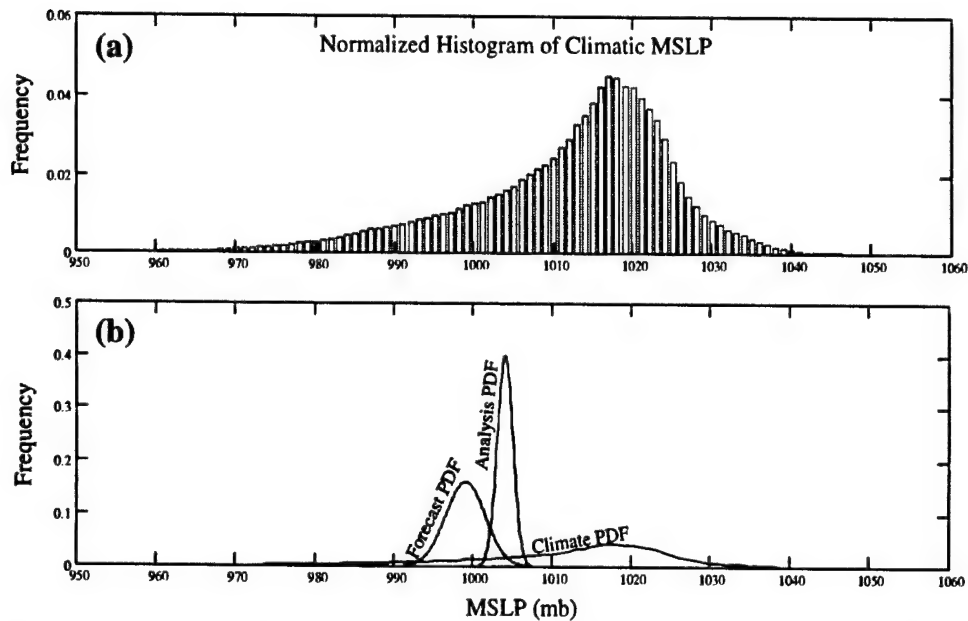


Figure 1. Example analysis, forecast, and climate PDFs for *MSLP*. (a) Histogram of the long-term observations. (b) Hypothetical analysis and forecast PDFs. The analysis PDF has an arbitrary  $\mu = 1004$  mb and observed average  $\sigma = 1.0$  mb. The forecast PDF has an arbitrary  $\mu = 999$  mb and observed average  $\sigma = 2.5$  mb for a 48-h forecast. The climate PDF is taken from (a) and has  $\mu = 1011.37$  mb and  $\sigma = 10.33$  mb.

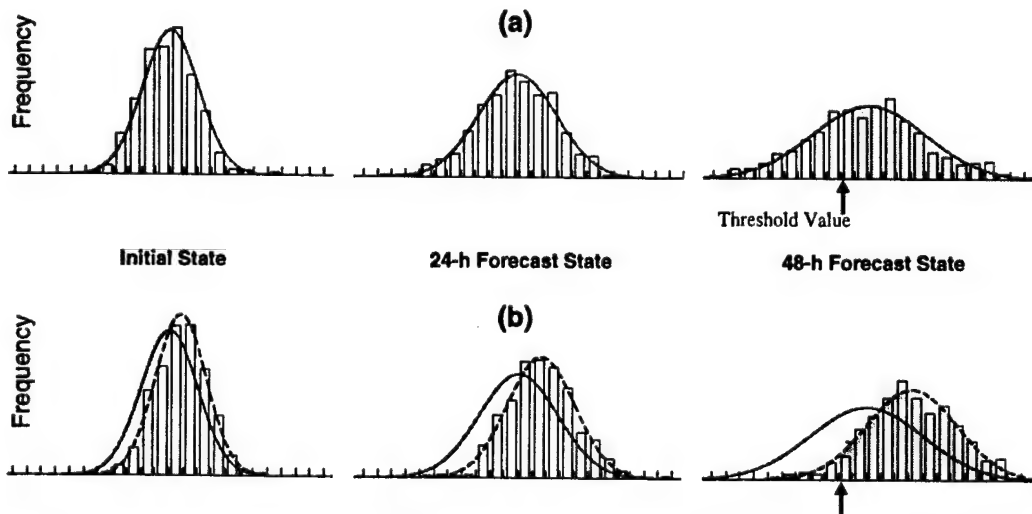


Figure 2. Simplified depiction of EF. The histograms represent a large ( $n = 500$ ) ensemble's estimation of the analysis PDF (solid curve at initial state) and forecast PDFs (solid curve at forecast states). The PDFs show possible states of the atmosphere, or simply the possible values of some parameter at a single point, such as surface temperature. (a) A well-calibrated ensemble which correctly estimates the PDF. (b) An inferior ensemble that incorrectly estimates the PDFs with a distribution (dashed curve) having a mean shifted to the right and too low a spread. The arrow is the event threshold, so the shaded region is the probability of exceeding that threshold.

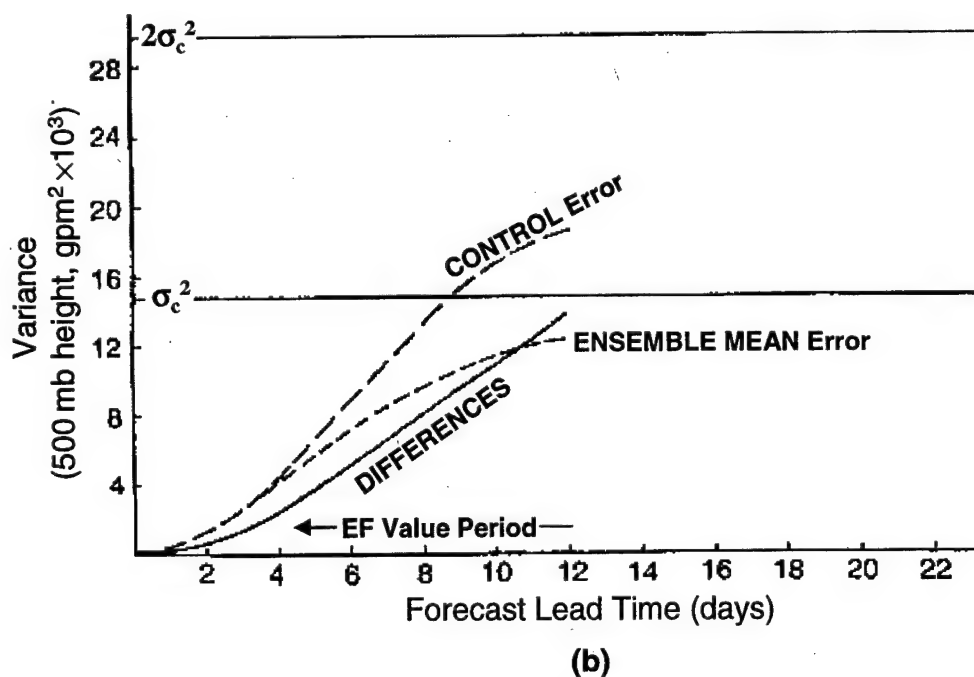
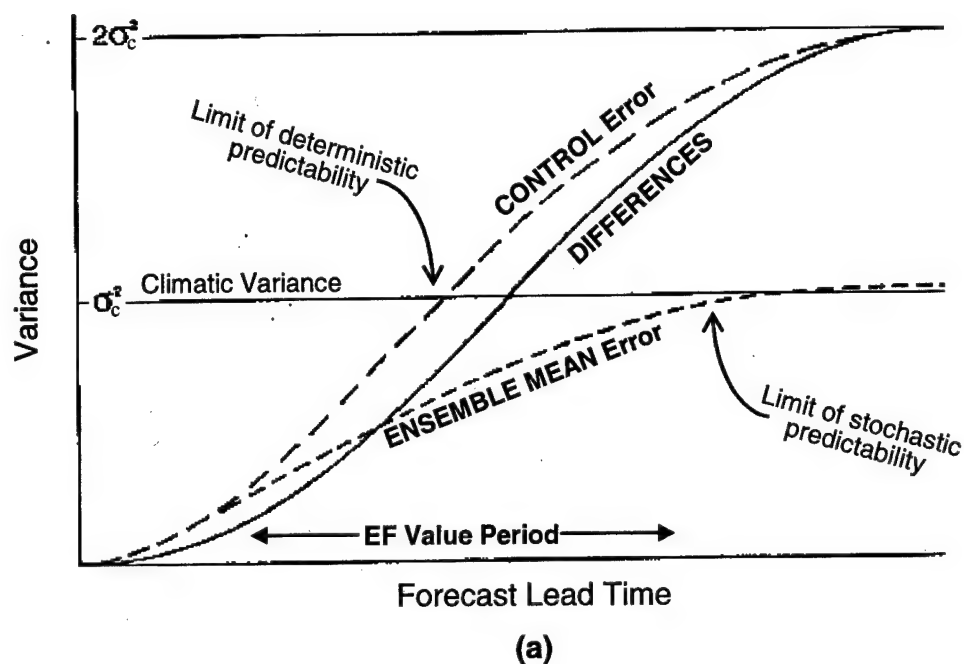


Figure 3. Error variance diagram examples (Baumhefner, 2000). (a) Illustration of the theoretical variances for a large sample of control forecast errors, ensemble mean forecast errors, and the ensemble differences as a function of forecast lead time. (b) Results for 45 ensemble forecasts of the 500 mb geopotential height field by the NCAR CCM3-T63 model, where forecasts were verified at one-day intervals over the region 130–70°W, 25–70°N.

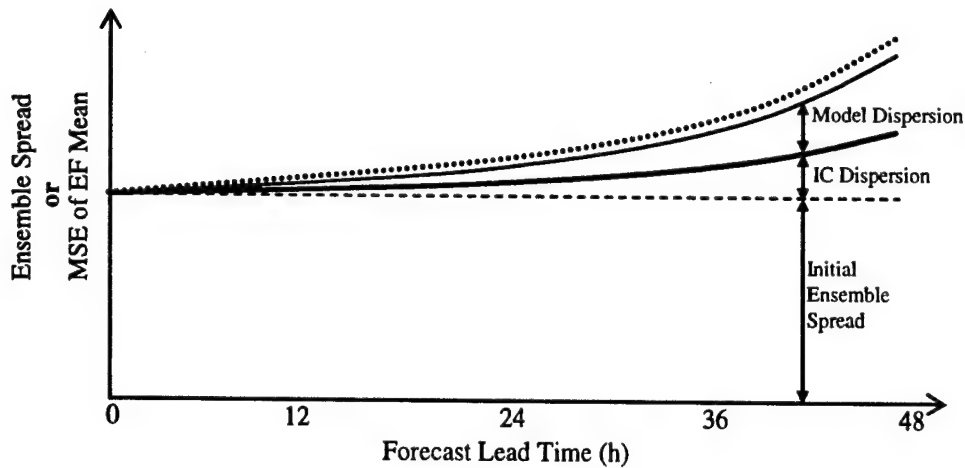


Figure 4. Dispersion diagram for a hypothetical 48-h forecasts of some variable. The dashed line shows the initial ensemble spread defined by spread among the set of ICs. The thick solid line shows the spread of an ensemble using those ICs and no model variations. The thin solid line is the ensemble spread for an ensemble using the same ICs and a unique model for each member. The dotted line is the *MSE* of the EF mean.

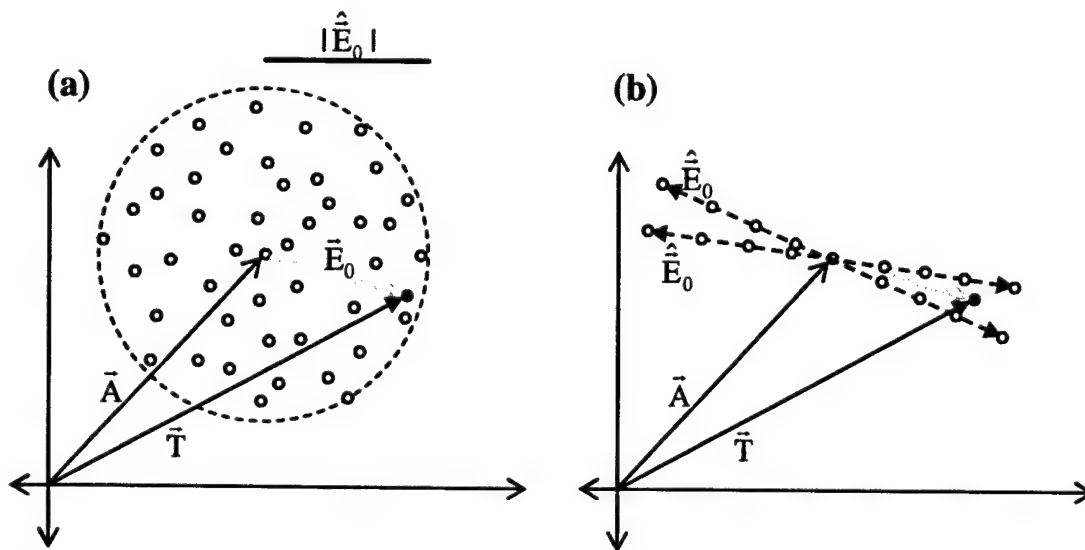


Figure 5. Simplified 2D generation of ICs from a best guess analysis  $\bar{A}$  given (a) only an estimate of the magnitude of analysis uncertainty and (b) estimates of the magnitude and direction.  $\bar{E}_0$  is the thick gray vector which points from  $\bar{A}$  to the true state  $\bar{T}$  in both panels.

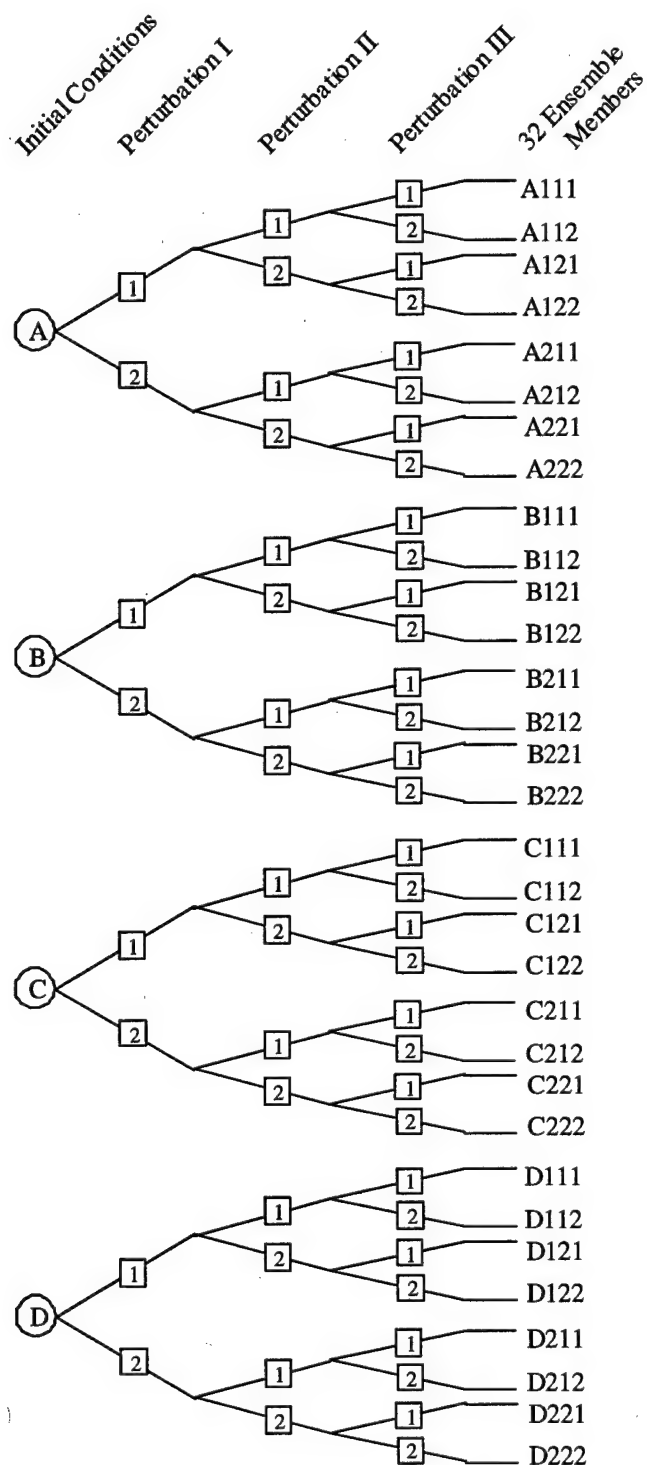


Figure 6. Tree diagram for an ensemble with 4 initial conditions (A, B, C, and D) and 3 model perturbations (I, II, and III) having 2 choices each. Provided that the model perturbations are uncorrelated and equally skillful, each of the 32 ensemble members yields an equally likely forecast.

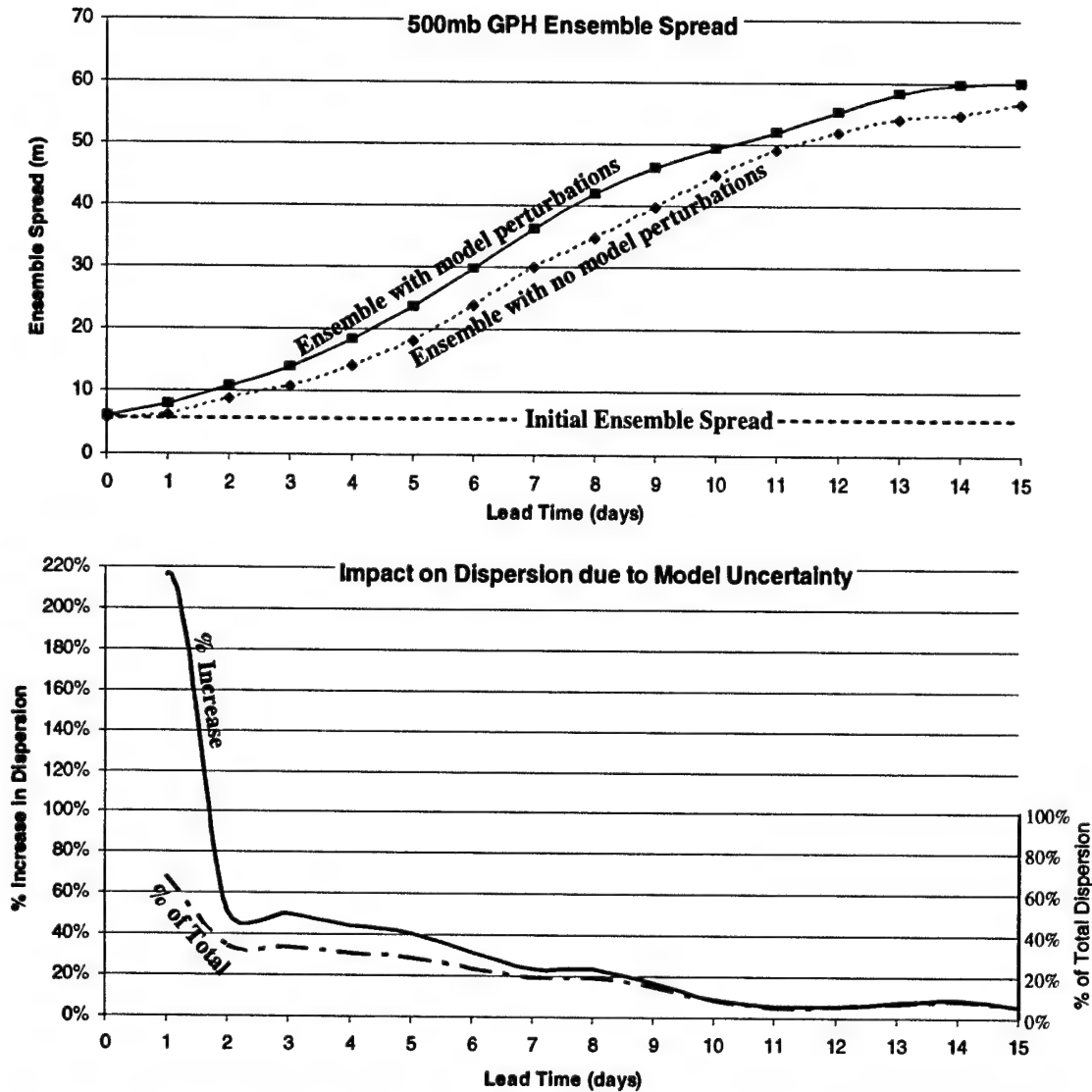


Figure 7. Impact of including model perturbations in a MREF. These plots were made from the data in Table 4 of Houtekamer (1996), which gives domain averaged ensemble standard deviation of 500 mb geopotential height (GPH) for a single EF case study. The upper plot is an empirical realization of Figure 4 but without the *MSE* of the EF mean and over much longer lead times. The lower plot shows how much model uncertainty may contribute to the dispersion.

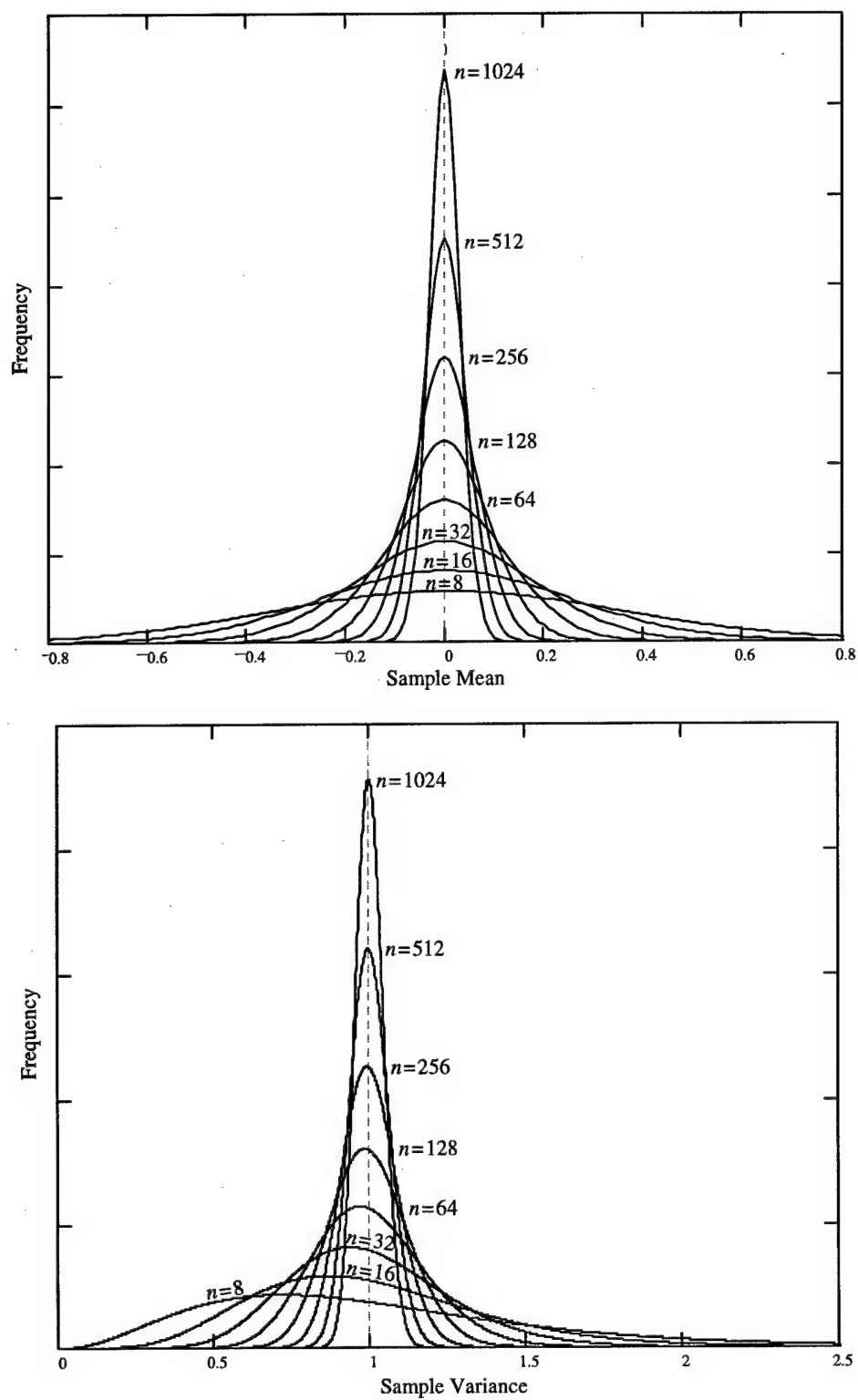


Figure 8. Sampling distributions of the sample mean and sample variance of an  $N(0,1)$  PDF. Each curve is for a different sample size  $n$ , labeled at or adjacent to its peak.

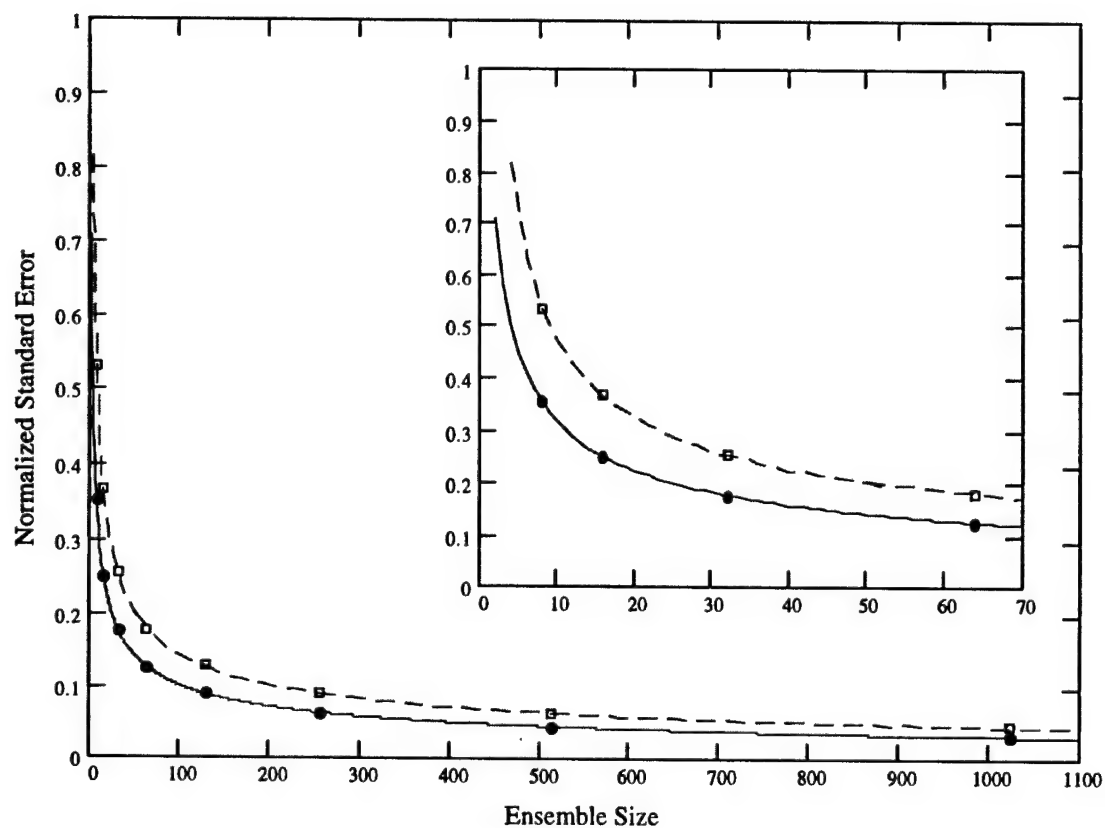


Figure 9. Standard errors of the sampling distributions for increasing sample size. The solid curve is the standard error of the sample mean as a fraction of the true  $\sigma$ . The dashed curve is the standard error of the sample variance as a fraction of the true  $\sigma^2$ . The standard errors from the curves for specific  $n$  values in Figure 8 are plotted as dots for the sample mean, and squares for the sample variance. The inset diagram is a zoom in of the region where increasing ensemble size yields the most benefit.



Table 1. Abridged list of three categorical sources of model error.

Model Error Category	Sources
Physics Parameterizations	<ul style="list-style-type: none"><li>- radiative transfer</li><li>- horizontal diffusion</li><li>- precipitation (droplet nucleation, growth, fallout, etc.)</li><li>- boundary layer behavior</li></ul>
Surface Boundary Parameters	<ul style="list-style-type: none"><li>- albedo</li><li>- roughness length</li><li>- ground temperature</li><li>- moisture availability</li><li>- sea surface temperature</li><li>- terrain height</li></ul>
Numerical Processing	<ul style="list-style-type: none"><li>- finite difference scheme truncation error</li><li>- precision of all variables and parameters</li><li>- internal precision of computer processor</li></ul>

## II. Methodology

In this chapter, we discuss the methodology applied in this SREF research. To test our hypotheses and assess the value of SREF, we designed a SREF test bed consisting of four distinct but related SREF systems (Table 2). The largest SREF system, Analysis-Centroid Mirroring Ensemble (ACME), was designed to improve SREF by using additional ICs. Our Poor Man's Ensemble (PME) is a collection of large-scale models run at different operational forecast centers. ACME<sup>core</sup> is our benchmark mesoscale SREF system that uses the PME's initial conditions (ICs) and lateral boundary conditions (LBCs), and a single version of MM5 for each member. ACME<sup>core+</sup> uses the same ICs/LBCs as ACME<sup>core</sup>, but each member uses a different (perturbed) version of MM5. These four systems have different strengths and weaknesses, and their intercomparison yields answers to the questions raised in Chapter I. The methods we employed in these systems for representing analysis and model uncertainty may be suboptimal but are functional enough to achieve our goals. Recall that the goal is to research fundamental aspects of SREF for the benefit of future systems and to design an effective SREF system with today's capabilities.

We are most interested in the cool season (Oct – Apr), when the Northern Hemisphere midlatitudes are prone to more rapidly changing synoptic conditions and thus when a SREF is likely to be of greater value. Additionally, SREF research to date has primarily focused on warm season data in which model uncertainty may play a greater role since weak synoptic forcing inhibits predictability error growth from analysis errors. In studying cool season data, we may gain more understanding of impacts to SREF from both analysis and model uncertainty. The 2001–2002 cool season was a test and development period for the ACME systems and is too incomplete for useful analysis. The probabilistic nature of EF requires a large number of cases to achieve statistical significance of results and reliable conclusions. During the 2002–2003 cool

season from 31Oct 2002 to 28 Mar 2003, we archived 129 forecast cases with complete data in all four SREF systems (Figure 10). This was a substantial accomplishment considering the complexity of the processing and extreme amount of data.

The common grid for the four SREF systems is a 36-km resolution domain depicted in Figure 11a. Imported model data of the PME was fit to the 36-km grid using bilinear interpolation programmed in the MM5 preprocessing code. ACME, ACME<sup>core</sup>, and ACME<sup>core+</sup> ran on the 36-km outer and 12-km inner MM5 domains (Figure 11b) using 32 sigma levels (31 layers). Forecasts from all configurations were initialized daily at 00Z and run through 48 h. The PME data was downloaded twice daily at 00Z and 12Z and archived at 6-h forecast intervals over the 48-h valid period (i.e., data at forecast hour 0, 6, 12, ..., 48). All ACME model runs were archived at 3-h intervals. Archived variables include winds at 10 m, maximum 3-h 10-m wind speed, moisture at 2 m, temperature at 2 m, maximum and minimum 3-h temperature, 3-h cumulative precipitation, and winds, temperature, moisture at the 850-, 700-, 500-, and 300-mb levels.

### A. Analysis Uncertainty

The mirroring approach used in ACME came out of a meeting with Dave Baumhefner in August of 2001. Basically, ACME expands upon the SREF research of Grit and Mass (2002) which showed that analyses from different operational centers provide practical ICs for a SREF. Their small, 5-member MM5 ensemble used multianalysis ICs (i.e., a set of five independent analyses) and successfully predicted forecast skill.

The reason why the multianalysis IC methodology works so well for a SREF over the Pacific NW is somewhat counterintuitive. It is logical to think that the IC perturbations for a short-range, mesoscale ensemble should include an estimate of errors on all scales with perhaps special attention to the mesoscale. Using global analyses as ICs provides little to no information

concerning mesoscale analysis errors since the differences among the analyses are predominantly on the synoptic scale. So how can these ICs be useful for a SREF?

Errico and Baumhefner (1987) showed that in general there is no need for ensemble ICs to include small-scale perturbations since predictability error growth is dominated by the synoptic scale. An ensemble containing ICs with only small-scale perturbations has extremely low dispersion while one with only large scale perturbations generates large dispersion on all scales. Using different analyses as ensemble ICs is therefore an excellent technique for SREF. One possible drawback, which will be discussed further below, is that the analyses may be too highly correlated to be considered random samples (Ebert, 2001).

Another reason for the success of using various analyses as the ICs is that, in the cool season, many mesoscale weather phenomena are driven by the synoptic-scale flow, particularly in areas of complex terrain such as the Pacific Northwest (Mass, 2002). For example, the position and intensity of the Puget Sound convergence zone is largely determined by the characteristics of the large-scale, low-level flow impinging on the Olympic Range. As a result, errors in the synoptic-scale flow cause the largest part of the forecast error within the Puget Sound. This example further supports the conclusion that the ensemble ICs should represent the likely errors on the synoptic scale, not small-scale errors.

By expanding the Grit and Mass (2002) SREF, ACME's objective is to provide an improved sampling of analysis uncertainty while maintaining the basic approach of using different analyses for ICs. From the original five analyses of Grit and Mass (2002), we first dropped the NCEP MRF model analysis since it is too highly correlated with the aviation (avn) model analysis. Next, we added four more analyses from different centers, bringing the total up to eight—collectively referred to as the *core* of the ACME ICs (Table 3). Eight random samples are likely still too few to thoroughly represent the analysis uncertainty, leading to an ensemble

forecast that may frequently poorly portray truth. We therefore attempted to use the ensemble core to generate more ICs, each with slightly different synoptic structures that are both realistic and within the bounds of uncertainty.

We began with the basic assumption that the core is a sufficiently diverse sampling to represent the general spread of the analysis PDF. Consider the core to be a rather sparse cloud of ICs that contains valuable information on analysis error. It seemed possible then to use the core to produce an estimate of the elusive analysis error vector, Equation (9), which would provide information on both error structure (i.e., direction in phase space) and error magnitude. Additional independent ICs could be created by varying the magnitude (i.e., changing the length of the error vector) within some predetermined bounds while maintaining direction (synoptic-scale structural information). Such a process would fill in the IC cloud and perhaps expand it, sampling likely ICs not represented in the core.

Our method to find  $\hat{\vec{E}}_0$  (the estimate of analysis error) begins with calculation of a *centroid* analysis,  $\vec{C}$ :

$$\vec{C} = \frac{1}{8} \sum_{i=1}^8 \vec{A}_i \quad (16)$$

which is the mean of the eight  $\vec{A}$  (core analysis) found by averaging all state variables, at all levels, over the entire model domain. This is considered our best estimate of  $\vec{T}$  (the true state) because it likely filters out the small-scale differences of the various  $\vec{A}$ 's that are likely to be in error (Richardson, 2001a). The centroid is run as yet another ensemble member and should on average be the most skillful deterministic run over a large domain, although it may occasionally be beat by another ensemble member because of the undersample problem discussed below.

Richardson (2001a) found that the centroid run from five independent analyses was slightly more skillful than the ECMWF out to seven days.

Since  $\bar{C}$  is our best estimate of  $\bar{T}$ , we substituted  $\bar{C}$  for  $\bar{T}$  in Equation (9) to get

$$\hat{\bar{E}}_0 = \bar{C} - \bar{A} \quad (17)$$

providing an estimate of  $\bar{E}_0$ . This  $\hat{\bar{E}}_0$  may be considered to be a perturbation to  $\bar{C}$  that produced  $\bar{A}$ . Such a perturbation could vary in magnitude or even reverse direction with respect to  $\bar{C}$  but still maintain structural error information, which is dominated by synoptic-scale errors.

Each of the eight analyses produces a different but somewhat correlated  $\hat{\bar{E}}_0$ . A new, valid IC ( $\bar{A}'$ ) could conceivably be placed anywhere along the two-way vector ( $\bar{C} - \bar{A}$  or  $\bar{A} - \bar{C}$ ) by simply adding  $\hat{\bar{E}}_0$  times some perturbation factor ( $\rho$ ) back onto the centroid:

$$\bar{A}' = \bar{C} + \rho \hat{\bar{E}}_0 \quad (18)$$

In theory then, there are an infinite number of new possible ICs for each of the eight analyses of the core.

However, testing revealed that most of the possible  $\rho$  values are not beneficial to our SREF. Finding the best  $\rho$  values to use turned out to be a trade-off between skill and dispersion. When we used a small  $\rho$  such as  $-1.0 < \rho < 1.0$ , we produced an  $\bar{A}'$  that was too similar to either  $\bar{C}$  and/or the parent  $\bar{A}$ . Thus no new information was gained running the forecast from  $\bar{A}'$  and the ensemble had weak dispersion. On the other hand, a larger  $\rho$  lowered the skill (in a *RMSE* sense) in the resulting forecast and created unrealistically large dispersion.

A logical choice then was to use only  $\rho = 1.0$  to produce forecasts with new, useful information and skill on a par with the forecasts from the core analyses. A new IC is then the

mirror of its parent analysis across the centroid. From each analysis we generate one more IC, giving us a total ensemble of 17 members. Combining Equations (17) and (18), the mirror ICs are generated by

$$\bar{A}' = \bar{C} + \rho(\bar{C} - \bar{A}) \quad (19)$$

However, using  $\rho = 1.0$  creates a problem in that the variance of the full ACME ICs is reduced compared to the core analyses, a statistical result of the small ensemble size. The sample variance of the core is

$$s_{core}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (20)$$

where  $n$  is the ensemble size,  $x_i$  is the value of member  $i$ , and  $\bar{x}$  is the sample mean, which is equivalent to the centroid. The variance of the ACME ICs is

$$s_{ACME}^2 = \frac{1}{2n} \sum_{i=1}^{2n} (x_i - \bar{x})^2 \quad (21)$$

where the sum goes to  $2n$  and not  $2n+1$  since the centroid contributes nothing to the sum. Note that the sample mean (i.e., the centroid) is the same for both. For large  $n$ , Equations (20) and (21) produce the same result, but for an  $n$  as small as eight the ACME ICs have a lower spread than the core analyses. We corrected for this by using a  $\rho$  designed to adjust the ACME IC's variance to match that of the core. (Note: Eric Gritit is to be credited for the following proof.)

To find the desired  $\rho$ , we begin by expanding Equation (21) as

$$s_{ACME}^2 = \frac{1}{2n} \sum_{i=1}^n [(x_i - \bar{x})^2 + (x_{n+i} - \bar{x})^2]$$

where  $x_{n+i}$  represents the mirrored values. We can then apply Equation (19) to the  $x_{n+i}$  term to get

$$\begin{aligned}
 s_{ACME}^2 &= \frac{1}{2n} \sum_{i=1}^n [(x_i - \bar{x})^2 + \rho^2 (x_i - \bar{x})^2] \\
 &= \frac{(1 + \rho^2)}{2n} \sum_{i=1}^n (x_i - \bar{x})^2
 \end{aligned}$$

Equating this result to Equation (20) we can solve for the desired  $\rho$ :

$$\begin{aligned}
 \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 &= s_{ACME}^2 = \frac{(1 + \rho^2)}{2n} \sum_{i=1}^n (x_i - \bar{x})^2 \\
 1 + \rho^2 &= \frac{2n}{n-1} \\
 \rho &= \sqrt{\frac{n+1}{n-1}} \quad (22)
 \end{aligned}$$

For  $n=8$ ,  $\rho=1.13$ , which is the perturbation factor we used for ACME.

One difficulty in analysis-centroid mirroring is in handling the state variable for moisture (relative humidity, *RH*). This problem exists because moisture varies over the interval bound by absolute dryness ( $RH = 0.0\%$ ) and saturation ( $RH = 100.0\%$ ). Mirroring of a large *RH* difference toward either boundary can produce an unphysical *RH* value. Other state variables such as pressure and temperature also have bounds, but the variable's range within the troposphere is rather limited, nowhere close to their bounds. (For example, *MSLP* has a physical boundary of 0.0 mb but it typically varies between 970 mb and 1030 mb. A mirrored value may end up being extreme but is always physical.)

The easiest way to deal with this problem would be to truncate the mirrored *RH* value at 0.0% and 100.0%. This however produces a mirrored IC with unrealistic moisture patterns, having large areas of dryness or saturation, and large moisture gradients. The alternative we employed is related to the Zeno's Paradox—the idea that you can never reach a wall since you keep going  $\frac{1}{2}$  of the distance toward it.



For example, refer to Figure 12a to see how we arrive at a perturbation value at some hypothetical grid point where the centroid moisture ( $RH_C$ ) is 80%. Given a moisture analysis value ( $RH_A$ ) of 60% (solid dot), the thick arrow shows the perturbation of  $RH_A$  to  $RH_C$ , which for this example is exactly half way toward complete saturation. The mirrored moisture value is then half way again toward saturation, yielding an  $RH_M$  of 90% (hollow dot). Figure 12 shows how we extend this technique so that the fraction

$$\frac{RH_A - RH_C}{|RH_A - boundary|}$$

is then the fraction of the remaining distance from  $RH_C$  to the boundary for arriving at  $RH_M$ . This makes the  $RH_M$  asymptote toward the boundaries as the perturbation increases and asymptote toward  $RH_M = RH_C + (RH_C - RH_A)$  as  $(RH_C - RH_A) \rightarrow 0$ , just like Equation (19). Notice that we use 10% as the lower boundary instead of 0% as this is an MM5 preprocessing requirement.

While this technique may appear at first glance as somewhat arbitrary, it actually makes physical sense. In fact, one could argue that the general mirroring technique does Zeno-type mirroring for all the state variables. It is simply not apparent since the perturbations in the other variables are so small compared to the distance to their boundaries that the mirroring has effectively asymptoted to Equation (19).

## 1. IC Strengths

There are many aspects of multianalysis and ACME ICs that should be most beneficial to our SREF systems. The primary strength is that they likely produce a reasonable sampling of analysis uncertainty from day to day for several reasons.

To begin with, the eight analyses are produced with different models at different resolutions, as well as variations in observation data and data processing. (Note that an interesting twist in this method is that there is actually an ingredient of model uncertainty used in defining the

analysis uncertainty.) The processing diversity results in a significant spread among the eight analyses, which subjectively appear to be different enough to be considered random samples of the theoretical analysis PDF. It is not clear whether the mirrored ICs may also be considered random samples, but they do contain unique, meaningful synoptic-scale differences because use of the centroid likely yields good estimates of  $\tilde{E}_0$ .

The key quality of multianalysis ICs is that the differences among analyses are predominantly on the synoptic scale—precisely what is desired for a set of ICs since the biggest error in an analysis is in phase and/or amplitude of synoptic weather systems. Furthermore, it is these large-scale errors that experience the most growth during the forecast integration as extratropical cyclones develop and propagate (Errico et al., 2002). A set of ensemble ICs should contain a spread of similar synoptic waves with slightly different phases and amplitudes, but within the bounds of analysis uncertainty. Whether or not the ICs also include small-scale perturbations may be irrelevant since those errors grow insignificantly or decay.

Figure 13 is a simplified demonstration of how ACME further samples the spread of synoptic waves. The solid lines, representing the core analyses of *MSLP* along a latitude line, give the general spread of synoptic waves which we then build on. The centroid, our best guess analysis, is in the middle of the eight analyses, as should be expected. Notice that the centroid does not get significantly biased in phase, amplitude, or frequency when compared to the averaged values of the individual analyses. Simplified experiments showed that the frequency of the centroid is slightly lower (~1%) compared to the average frequency of the core analyses and that the amplitude of the centroid can be lower by up to a few percent.

A mirrored IC was simulated in Figure 13 by taking the difference between the centroid and one of the core analyses then projecting the reverse of that difference onto the centroid, as in Equation (17). The new IC therefore contains synoptic-scale error information from  $\hat{\tilde{E}}_0$ , resulting

in a different possible large-scale wave. Notice that the new wave is unique but subjectively looks like it could be just another one of the core analyses.

Another strength of using multianalysis ICs is the low computational requirement in the preprocessing phase, which includes downloading data, fitting data to the MM5 grid, and establishing the LBCs of the limited area domain (Figure 11). Each of the eight analyses can be downloaded and run through the MM5 preprocessing within a matter of minutes. Their LBCs are set by the forecast grids from the original model run. (E.g., the MM5 run using the avn analysis IC uses the avn's original forecast grids to define the MM5 LBCs.)

The additional preprocessing for the mirrored ICs of ACME is fairly straightforward and also computationally affordable. The mirrored LBCs follow exactly the same perturbation method as the mirrored IC fields. (E.g., for the MM5 run using the avn mirrored IC, the LBCs at the 6-h forecast point are the mirror of the 6-h avn LBCs across the 6-h centroid LBCs.) Additionally, the mirrored ICs are dynamically balanced on the large scale so no special processing is required. At small scales there are likely significant imbalances in the mirrored ICs since the core analyses themselves are not balanced with respect to the scales represented within MM5. MM5 handles this problem with strong diffusion, quickly damping out gravity waves.

## 2. IC Deficiencies

There are several possible deficiencies in the basic design of multianalysis and ACME's ICs. One problem is the low sample size. Considering the extremely high number of dimensions of the atmosphere, the 17 ICs of ACME or the 8 ICs of ACME<sup>core</sup> or PME are likely too few to consistently produce a reasonable representation of the analysis PDF, regardless of how ideal these ensemble systems may or may not be. A second problem is that the analyses may be too highly correlated and not independent, random samples of the analysis PDF (Ebert, 2001). This second problem would result in limited spread among the analyses (i.e., an analysis PDF with low

variance). Thirdly, the analyses (and resulting forecasts) are not equally likely, thus violating one of the basic tenets of EF. The combination of these deficiencies could seriously undermine our quest for an effective SREF.

To explore how the undersampling problem impacts our  $n = 8$  ensembles (PME, ACME<sup>core</sup>, and ACME<sup>core+</sup>), we temporarily ignored the second potential problem by assuming that the analyses are totally uncorrelated, random samples, and that differences between ICs truly represent analysis errors. An infinite number of these analyses would provide a perfect and complete analysis PDF, from which truth would always be a random sample.

Back in section I.B.4, it was shown that sampling with only a few random draws makes it difficult to recreate the PDF from which samples are drawn. Even when ensemble members are drawn from the same PDF as truth, as they should be, the EF estimate of the forecast PDF will often be in error and sometimes severely so. Continuing the sampling experiment introduced in section I.B.4., eight samples of a random variable  $x$  with a set PDF were taken repeatedly. For a more realistic simulation of EF, the random variable was chosen as 48-h 500 mb height at some grid point, drawn from a normally distributed forecast PDF with  $\mu = 5400$  gpm and  $\sigma = 15$  gpm, a typical forecast error. Over many trials then, an ensemble of eight members attempted to represent that forecast PDF. We then observed the behavior of the sample mean  $\bar{x}$  and the sample variance  $s^2$  to understand how their errors may affect the skill of our ensembles.

In Figure 14 (data values provided in Table 5) three example attempts to represent the forecast PDF are shown to demonstrate that with only eight members, it is easy to misrepresent the forecast PDF. Too high a spread (as depicted in Figure 14c) is not as significant a concern because, while it may be misleading for uncertainty, it still may reveal the different forecast possibilities and portray the true future state. Of more concern is too low a spread (as depicted in Figure 14a) where uncertainty is underrepresented and potentially important parts of the PDF go

unsampled. Equation (19) was applied to the data and displayed in the right hand side panels in Figure 14 to demonstrate how the generation of the mirrored ICs can ameliorate the undersampling problem by filling in the distribution and sampling a slightly wider region. The mirrored ICs should provide realistic, independent samples for a more complete representation of the analysis PDF.

After repeating 5000 realizations such as those in Figure 14, we plotted Figure 15 and Figure 16, which are repeats of the  $n = 8$  curves from Figure 8, but include the experimental, histogrammed sampling distributions from this simulation. It is evident that the high variance of  $\bar{x}$  and  $s^2$ , due to the small sample size, causes these sample statistics to frequently have significant error, producing a poor estimate of the forecast PDF. For example, the spread in Figure 14a, which is noticeably too low, is not an extreme value of the sampling distribution of  $s^2$ .

One question that arises is: which error causes more problems for the EF, incorrect location (ensemble mean) or incorrect spread (ensemble variance)? We can address this from the point of view of the *FP* derived from an estimated forecast PDF. Using the results of Figure 9 and Equations (14) and (15), the magnitude of the standard error in  $\bar{x}$  is 5.3 gpm, and 120.1 gpm<sup>2</sup> for  $s^2$ . Using these values, the erred distributions along with the correct PDF are plotted in Figure 17a, and the PDF of their combined effect is plotted in Figure 17b. (Note that we chose the positive  $\bar{x}$  deviation, giving  $\bar{x} = 5405.3$  gpm, and a negative  $s^2$  deviation for  $s^2 = 104.9$  gpm<sup>2</sup>.) For any given event threshold value of 500 mb height, each PDF yields a different value of *FP* (area under the curve to right or left of the threshold). The exception is when the event threshold falls beyond about  $3\sigma$  when each PDF yields an *FP* of 0.0 or 1.0. Figure 17c and d show *FP* for the full range of event thresholds where the probability of exceeding the event threshold is forecast. Plotting the *FP* error (correct – erred) in Figure 17e reveals that the standard error in the

mean actually causes larger error in *FP* than that of the standard error in the variance, thus highlighting the importance of bias correction.

The summary plot in Figure 17f shows how the combined effect of typical mean and spread errors impact an  $n=8$  ensemble's *FP*. Low sampling causes significant errors in the midrange *FP* when the event threshold falls within about  $1.0\sigma$  of the governing PDF. Note that this effect can not be calibrated out of the system because it is totally random. It is something we must live with and consider when analyzing our results.

Let us now consider the second potential problem of correlation among the analyses. In any ensemble system, the set of ICs will naturally be somewhat correlated since they are all attempting to describe the same instantaneous state of the atmosphere. It may be that, for multianalysis ICs, the level of correlation is too high because the analyses are built using comparable observational data, making them share similar errors.

If we assume some high level of correlation among the eight samples in the above simulation, the ability to reasonably represent the PDF worsens. A strong correlation between the analyses would reduce the sample variance, limiting the ensemble's ability to portray the true state. Also, if the analyses share similar biases, the error in  $\mu$  would increase. Ebert (2001) showed that the correlation of precipitation forecasts among the members of a PME is acceptably low. This is encouraging but it is unclear if it holds true for the state variables of the analyses.

The third potential deficiency—lack of equal skill among the analyses—is like supposing that the analyses are drawn from separate PDFs in which a less skilled analysis is associated with a wider PDF. In that case, the ensemble's PDF may be meaningless. However, since each of the different analyses' PDFs may be a fair estimate of the true PDF, the ensemble's PDF may contain a good representation of analysis uncertainty. We will explore this issue further below when it is additionally complicated by the use of different models. For now, we simply note that one source

of the inequality in the solutions of our SREF systems is the different levels of skill in the analyses.

As a final note in this section, there are some significant technical challenges for a multianalysis ensemble. Besides the problems of undersampling, such a system is at the mercy of the analyses in other ways too. There is a delay in downloading all the analyses, reducing the utility in running the SREF in real time. The system is also apt to occasionally miss analysis data since so many data sources are relied upon. Lastly, frequent updates in the techniques employed at the operational centers to produce the analyses affects our ability to design a calibration based on identifying and correcting for systematic errors (Eckel and Walters, 1998). Alterations to the source model or objective analysis scheme invalidate a calibration based on the former analysis.

## **B. Model Uncertainty**

This section discusses the methodology of two techniques—perturbed-model and multi-model—for representing model uncertainty that we employed in this research. Since a single model EF system is generally found to be underdispersive, the goal of including model diversity in an EF is to increase dispersion. This should produce a more accurate estimation of the forecast PDF and thus more highly skilled forecast probability (*FP*).

One issue common to both techniques is the lack of equal skill among the members. We noted above that since there is inequality among the core analyses, we can expect single-model SREF systems to have solutions that are not equally likely. As we attempt to account for model uncertainty by varying the model, we are likely to make the relative skill among the members even more disparate.

Mylne (2002) indicated that having unequal members is not problematic and may in fact be advantageous. Expanding an EF by including inferior models can be beneficial to a SREF system because of the added diversity. Evans (2000) points out that the key to the benefit of a MMMA

system is for the models to sample different, plausible regions in phase space, where the true state may lie. Models having different strengths and weaknesses can be combined to make a system that outperforms an EF that uses only one model. A model may be inferior overall but still add some skillful information to the ensemble if it occasionally performs better at some locations or with some phenomena.

Use of unequally skilled members is an apparent failure to meet one of the fundamental objectives of EF. Members that are not equally likely can not be considered independent, random draws from the same forecast PDF. In fact, when different models are involved, each member is really drawn from a different PDF since each model has its own attractor. In a model with higher skill, error growth is slower, so its solution at some lead time (before error saturation) is drawn from a relatively narrow PDF. Likewise, a lesser skilled model solution comes from a wider PDF. The ensemble forecast PDF is actually an amalgamation of samples from many different PDFs. It is precisely this mixing of information that results in accounting of model uncertainty by either the perturbed- or multi-model approach.

### **1. Perturbed-Model Application**

ACME<sup>core+</sup> (see Table 2) applies the perturbed-model strategy by using the same ICs as ACME<sup>core</sup> and a uniquely perturbed version of MM5 for each of the eight members. As previously discussed, representing model uncertainty with the perturbed-model strategy is potentially rewarding but difficult to apply. The variety and number of model error sources make it nearly impossible to completely and accurately represent all model errors in a SREF. The methodology employed in ACME<sup>core+</sup> is meant to capture a significant portion of the model error in order to explore the benefits and potentially to realize an effective SREF.

The focus of ACME<sup>core+</sup> was not to improve deterministic MM5 forecasts but rather to represent the uncertainty present in MM5. We perturbed as much diversity as possible in order to



generate large and realistic dispersion, thus losing the ability to ascertain an optimally perturbed deterministic model configuration. (I.e., we were not running as system simulation experiment, SSE.) Furthermore, all perturbations were made in keeping with the original MM5 design. That is, we did not use experimental perturbations designed to improve the model, but rather made perturbations that preserved the original design of the MM5 routines.

ACME<sup>core+</sup> does have some similarities to a SSE since each model version is a fixed combination of model options and perturbed surface boundary parameters (SBPs). Referring back to Figure 6, ACME<sup>core+</sup> consists of a fixed set of branches where each branch begins at a different IC. The difference from a true SSE is that each branch is designed to be as unique as possible. A SSE tries to determine an optimal model set up by limiting model option combinations.

The major factors that were considered in designing the MM5 model variations (i.e., building the branches) were:

- 1) **Sensitivity.** Since generating increased, useful dispersion was the main objective, the primary consideration in choosing model aspects to perturb was their sensitivity. We sought to alter anything that made a large difference in the solution when perturbed within its suspected uncertainty.
- 2) **Uncertainty.** Another critical consideration was to perturb model aspects that contain large uncertainty. A parameterization that shows large sensitivity but is well known or well represented may not be worthwhile to perturb. This is likewise for a parameterization that has large uncertainty but little sensitivity.
- 3) **Feasibility.** The final consideration was that the model aspect should be fairly easy to alter within the MM5 model. For example, the forecast is certainly sensitive to the

numerical methods within MM5, but altering them would involve a major rewrite of the MM5 code. Such a perturbation is beyond the scope of this research.

Table 4 lists the eight branches chosen for ACME<sup>core+</sup>. By Equation (10) there are actually 1,228,800 possible branches to choose from. We only used eight of these since our objective is not to span the space of model uncertainty but simply to represent model uncertainty in a SREF. One thing to note is that the focus of these model perturbations is on the solution at or near the surface. This was not originally intended as part of the design but came about naturally as perturbations were selected because model parameterizations cause the greatest error at the surface and lower atmosphere (Stensrud et al., 2000).

Table 4 also shows the MM5 version shared by all members of ACME and ACME<sup>core</sup>, the single-model SREF systems. Over the course of many previous studies at the University of Washington, these are the model options determined to perform the best over the Pacific Northwest and are therefore used in the high-resolution deterministic forecast system. It is therefore expected that the MM5 versions of ACME<sup>core+</sup> should exhibit less skill compared to the parallel component forecasts of ACME<sup>core</sup>. But as discussed above, a member may add valuable information to an ensemble if it can occasionally perform better. Figure 18 shows an example forecast verification comparison between the eta member of ACME<sup>core</sup>, and the plus03 member of ACME<sup>core+</sup>. Plus03 was able to outperform the eta over significant regions, showing that it is a valuable EF member.

One last thing to note from Table 4 is that not only are the model variations held constant, but they also remain tied to a particular IC. One could argue that this severely constrains the SREF system since applying more randomness among the variations from day to day would capture much more of the possible model error over many case days. We opted to fix the system because of that fact that the members have unequal skill. With a fixed system, we have a chance to

remove bias and possibly produce calibrated probabilities. (A bonus is that it is also much easier to program.) Each member of ACME<sup>core+</sup> likely has unique biases, coming from both the IC and model version. To produce more skillful probability forecasts, this bias should be removed. By keeping the IC and model version for each ACME<sup>core+</sup> member fixed, we can determine bias from a record of previous forecasts and observations.

#### a) Model Options

An important question for the perturbed-model method concerns whether the differences between model options really represent model uncertainty. For example, consider two values for cumulative precipitation produced by the Goddard and the Shultz precipitation schemes. Does the difference in the two values reflect either scheme's (or the model's) inability to accurately represent the precipitation process? Or are the two schemes both so oversimplified and parameterized that the difference between them is meaningless? Unfortunately, these questions can not easily be answered. For this research we made the large and potentially harmful assumption that differences between model options are reasonable approximations of model uncertainty. The solution from different schemes can often be dramatic because they may not simply be using different values of some parameter but also a completely different methodology of modeling a physical phenomenon.

We were able to generate considerable diversity among MM5 solutions by choosing various combinations of model options for each ensemble member, which given our assumptions means that this diversity represented much of the likely model uncertainty. In order to get the most variety, the MM5 versions shown in Table 4 were set up to be as different as possible, but some limitations were imposed by the design of the MM5 code. For example, the land surface model (LSM) code is only compatible with the MRF and Eta PBL schemes.

The Reisner II, Skip 4 cloud microphysics scheme is a modified version of the standard Reisner II scheme. To speed up this extremely costly code, production terms are held constant for four time steps. Garvert (2002) found that this does not change the solution appreciably but decreases total run time by about 1/3.

#### b) Perturbations to Surface Boundary Parameters

Accounting for the uncertainty in a SBP is accomplished in a more idealized sense by designing random perturbations to mimic the suspected errors. This is much different than applying different model options where we hope that differences represent model uncertainty. When perturbing a parameter directly, we have much more flexibility. The reason all model aspects were not similarly handled is that these errors are often poorly understood and/or extremely difficult to directly perturb. We generally have some idea of the errors in SBPs, and they are also fairly straightforward to perturb. In this section we will describe how our perturbation methodology for sea surface temperature (SST), moisture availability, albedo, and roughness length is designed to provide a reasonable representation of model errors from these sources. Note that even though we have more flexibility in designing SBP perturbations, we chose to keep them fixed once constructed since varying the SBP perturbations randomly from day to day would likely have reduced the effectiveness of the bias correction.

The four SBPs we chose to perturb were selected because they strongly satisfied the design considerations of sensitivity, uncertainty, and feasibility. The model solution is quite sensitive to small changes in these SBPs, they have a significant amount of uncertainty in their value, and they are fairly easy to alter. Their most significant direct impact is to the surface energy equation:

$$C_g \frac{\partial T_g}{\partial t} = R_n - H - G - L_v E \quad (23)$$

where  $C_g$  is the slab thermal capacity [ $\text{J m}^{-2} \text{K}^{-1}$ ],  $T_g$  is ground temperature [K],  $R_n$  is the net radiation at the surface [ $\text{W m}^{-2}$ ],  $G$  is the heat flux into the substrate [ $\text{W m}^{-2}$ ],  $H$  is the sensible heat flux into the atmosphere [ $\text{W m}^{-2}$ ],  $L_v$  is the latent heat of evaporation ( $2.5 \times 10^6 \text{ J kg}^{-1}$ ), and  $E$  is the evaporation rate at the surface [ $\text{kg m}^{-2} \text{s}^{-1}$ ]. This equation is used to estimate the tendency in  $T_g$ , a major component in the behavior of the planetary boundary layer (PBL). Indirectly then, our perturbations significantly affect phenomena such as lower atmosphere air temperature, stability, surface winds, cloud height, and precipitation. Indeed, these are exactly the phenomena for which we wish our SREF to represent the full range of possible values. Note that our model option variations are also directly or indirectly affecting Equation (23). For example,  $R_n$  is altered in our variation of the radiation scheme.

The four SBPs affect Equation (23) and other model aspects in various, complex ways. To make accurate perturbations for these SBPs, and any model aspect for that matter, it is desirable to thoroughly define their uncertainty. To demonstrate how difficult such an investigation is, Appendix II includes a lengthy review of how MM5 models evaporation rate with the moisture availability SBP. Even after that investigation, one is left with only a vague idea of the uncertainty involved with moisture availability. Therefore, in this dissertation we avoid a lengthy discussion of exactly how the SBPs are modeled, their various direct and indirect effects, and implications of their uncertainties since such discussion is not productive to our goals.

The difficulty in quantifying model uncertainty for constructing model perturbations is a big problem we faced in designing ACME<sup>core+</sup> and a general problem that EF will likely always have. The best we can do is to use completely different modeling approaches by selecting different model options, and make reasonable approximations for the uncertainties in SBPs. In the end, if the ensemble with the increased dispersion from model diversity performs better, then we can conclude that the methodology was at least sound.

Disregarding SST, one aspect of the uncertainty in the SBPs that we could quantify somewhat came from the manner in which the SBP values are employed. A single value for each SBP for a certain model grid box is taken from a look-up table (Table 6) by having each model grid box assigned one of 24 *land use* values. The SBP values in the MM5 land use table were designed to produce long-term average results that agree with climatology, which means the SBP values can be significantly in error on any particular forecast cycle (Bretherton, 2002). The use of fixed values in a grid box further increases the uncertainty since the land use identification is simply determined by the dominant type of surface present. For example, Figure 19 shows that the 36-km grid boxes over the Puget Sound are all considered to be evergreen needleleaf forest, although most contain a significant amount of open water. Therefore the model will likely underestimate the evaporation rate in these grid boxes by applying too low a value of moisture availability.

To account for the uncertainty in moisture availability, albedo, and roughness length, we designed a unique PDF for each SBP at each land use to represent the possible values of the SBPs. This process involved a combination of empirical evidence, logic, conjecture, and a good deal of imagination. All PDFs were based on the gamma PDF, Equation (24), because of its ability to take on a wide variety of shapes.

$$f(x; \alpha, \beta) = \frac{1}{\beta^\alpha \Gamma} x^{\alpha-1} \exp(-x) \quad (24)$$

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$$

where  $x$  is the random variable (albedo, moisture availability, or roughness length),  $\alpha$  is the shape variable, and  $\beta$  is the spread variable. For additional flexibility, we added two more variables:

$$f(x; \rho, \alpha, \beta, \zeta) = \frac{1}{\beta^\alpha \Gamma} \left( \frac{\rho(x-\zeta)}{\beta} \right)^{\alpha-1} \exp\left( \frac{-\rho(x-\zeta)}{\beta} \right) \quad (25)$$

where  $\rho$  is a reversing variable and  $\zeta$  is a translation variable. Thus we needed to define four adjustable variables ( $\rho$ ,  $\alpha$ ,  $\beta$ , and  $\zeta$ ) to define each unique PDF using Equation (25).

The only concrete evidence we had concerning the possible range of values for the SBPs was the two seasonal (summer vs. winter) values in the standard land use table and empirical data from tables 7-2 and 11-4 in Pielke (2002). This gave us a general idea of how much uncertainty (i.e., variance) to build into each PDF. Additional variance was included to account for the limitations in the gridded land use process. The values of the 576 required gamma variables (4 variables for 3 different SBPs with 24 land uses and 2 seasons each) of the PDFs are listed in Appendix II. A few example PDFs are shown in Figure 20.

Once all the PDFs were defined, we produced eight new land use tables (listed in Appendix II), one for each member of ACME<sup>core+</sup>. The process involved generation of a random deviate from each PDF as perturbed values of the SBPs. Assuming our PDFs represent the uncertainty of the parameters, each resulting land use table is as valid as the original standard. One limitation of this method is that in using a unique but fixed land use table for each ensemble member, we restricted the diversity of the perturbations to being uniform in space as well as in time. That is, the same perturbed value for a particular parameter and land use is applied throughout the domain, rather than a different perturbation at every grid box with that land use. This was done to satisfy our strategy of preserving the basic MM5 modeling structure where every grid box with the same land use uses the same parameter values.

SST is modeled much differently in MM5 compared to the other three SBPs, so our perturbation technique is different. During the preprocessing phase, MM5 ingests a SST analysis field (for example see Figure 21) produced at 12Z daily by the Fleet Numerical Meteorology and Oceanography Center (FNMOC) with the Optimum Thermal Interpolation System (OTIS, described by Clancy and Sadler, 1992). The  $0.2^\circ \times 0.2^\circ$  data is fit to the MM5 grids with bilinear

interpolation and then held constant during MM5 forecast integration where it is used to determine the heat and moisture fluxes over water.

Given the significant influence that the eastern Pacific has on our forecast region, small SST errors over large areas may result in notable forecast errors. Suppose SST was analyzed  $1.0^{\circ}\text{C}$  too low over a large region where an extratropical cyclone is developing before moving on shore. The model's surface evaporation rate would be slightly too low, which, given time, would result in reduced moisture well up into the atmosphere. This would lead to a cascade of further effects but, most notably, reduced precipitation when the storm makes landfall.

Holding SST constant during the 48-h forecast period introduces a small error. In the open ocean, the diurnal variation of SST is on the average  $0.2^{\circ}\text{C}$  to  $0.3^{\circ}\text{C}$  (Clancy and Sadler, 1992). This can be much higher near land or if the water is suddenly well mixed. The more significant error comes from the SST analysis cycle (a process similar to an atmospheric analysis cycle) where an objective analysis routine combines buoy and satellite observations with an OTIS model first guess. Clancy and Sadler (1992) suggest that the typical SST *RMSE* is  $0.5^{\circ}\text{C}$ – $1.0^{\circ}\text{C}$  in our domain. Furthermore, the errors have a high degree of spatial correlation with a length scale of roughly 150 km (Cummings, 2002).

We attempted to design SST perturbations to mimic the likely error field. This was done by seeding a small field with random numbers, which were then smoothed and stretched to produce a field covering our domain and having coherent structure and a somewhat conservative average perturbation of  $0.7^{\circ}\text{C}$ . As an example end result, Figure 22 shows the perturbation for member plus01 of ACME<sup>core+</sup>. The inner domain's perturbation was made to match up with the outer domain to maintain consistency. Figure 21b is the resulting SST analysis when the perturbation is applied to the original SST analysis (Figure 21a). See Appendix II for all eight SST perturbation fields.



## 2. Multimodel Application

Application of the multimodel strategy is quite straightforward compared to the perturbed-model strategy. The challenge of deciding how to represent the various sources of model uncertainty is accomplished by simply using many completely different models in the ensemble. Each model may contain unique physics packages, SBPs, numerics, resolution, boundary techniques, and vertical coordinate. In general then, we should expect much more diversity among members of a MMMA ensemble versus those of a PMMA ensemble, and thus greater dispersion as well.

Whether or not the greater dispersion of the multimodel method is a more complete representation of uncertainty has to be determined. Just as with PMMA there is still the question of whether the model differences between MMMA ensemble members are representative of model uncertainty. When different models share similar limiting assumptions, the difference in their solutions would underestimate model error. It is also possible that drastically different models produce such dissimilar results (perhaps with oppositely signed errors) that their differences could overestimate the error of either model. In that case, a MMMA ensemble would be overdispersive, producing a forecast PDF with too much variance.

Our MMMA ensemble system is the PME, a group of independent, operational, large-scale models (Table 3). With our focus on the mesoscale, the original purpose of importing all these data was to provide the ICs and LBCs for the ACME systems. We soon realized however that there is value on the synoptic scale in considering these original forecasts as a separate, complete EF system. While the PME suffers from a much lower resolution, it may benefit from the greater diversity generated by the differences in the eight models.

By comparing the PME with ACME<sup>core</sup> we can explore how the increased dispersion provided by the multimodel technique affects ensemble performance. Furthermore, by

comparing the PME with ACME<sup>core+</sup> we can assess the differences between the PMMA and MMMA methods in accounting for model uncertainty. Model resolution becomes a serious issue when performing this later comparison. Comparing the skill of a set of mesoscale models to that of a set of global models is sketchy at best. We will attempt to account for this in our analysis by evaluating only large scale features, using the outer 36-km domain.

### C. Postprocessing and Analysis

In this section we will discuss how the SREF data were processed after completion of the MM5 runs. We begin with a description of the data used as verification, which has large implications for post processing and analysis. The two major postprocessing steps described here are bias removal and calculation of *FP*.

#### 1. Verification

A variety of statistical tools and metrics were used to evaluate and compare the skill of the four SREF systems, including standard deterministic error measurements such as root-mean-square error (*RMSE*) plus statistical tools tailored specifically to measuring EF skill such as the verification rank histogram, reliability diagram, Brier skill score, and relative operating characteristic diagram (see Appendix I).

The most critical question in any type of model verification is what to choose as truth. Since the true state of the atmosphere can never be known precisely, there exist many different approximations for it. The characteristics of the approximation employed in verification must be considered since this can significantly influence the results.

The primary type of truth used as verification was model-based, gridded analysis. The big advantage is that it provides complete coverage (both horizontally and vertically) over the model domain so we can generate a large sample of forecast/observation data pairs. A large sample is

absolutely essential when assessing the quality of an EF system because of the probabilistic nature of EF. The disadvantages of verifying with a gridded analysis are: 1) the value is dependent on the quality of the analysis (e.g., biases in the analysis from use of model first guess can lead to poor error estimation); and 2) the scales resolved by the analysis must be compatible with those of the forecast.

For the outer 36-km domain, we chose to use the centroid analysis as verification, rather than one of the eight PME analyses, as truth since the centroid analysis likely contains the least amount of error and bias. The verification centroid analysis is slightly different from the centroid analysis used as an IC in ACME in that we omitted one of the analyses from the Taiwan Central Weather Bureau since it proved to contain much more error compared to the other 7 analyses. Also, because of its low resolution, the centroid analysis is not appropriate for verifying the mesoscale forecast information of our inner domain. In fact, it does not even contain many variables of interest, such as temperature at 2-m ( $T_2$ ) and 10-m wind speed ( $WS_{10}$ ). Therefore we chose to use the mesoscale analysis provided by the Rapid Update Cycle 20-km resolution modeling system (RUC20, Benjamin et al., 2002). The RUC20 produces a new analysis every hour using a 3-Dimensional Variational Data Assimilation (3D-Var) scheme to combine a first guess from its 50-level mesoscale model with a large variety of observational assets.

To make the RUC20 analysis a fair verification of the 12-km MM5 data, the 12-km forecast data was smoothed out to the RUC20 grid using bilinear interpolation. Figure 23 gives a sample result of this refitting process and a RUC20 analysis. The grid alignment is different between the two grids so the 12-km data appears skewed within the 20-km domain. Wind barbs are at every 5<sup>th</sup> grid point on the 12-km plot (60 km apart) and every 3<sup>rd</sup> grid point on the 20-km plots (60 km apart). The smoothing of the isopleths from the 12-km data to the 20-km data is most evident over land, but the solution remains essentially the same. It is unclear why the RUC20 analysis

*MSLP* appears smoother compared to the 20-km MM5 forecast. Additionally, the RUC20 analysis has notably more variance in its  $T_2$  analysis, and subjectively appears to be a more appropriate representation of mesoscale features compared to the MM5 forecast.

To confirm some of our results, we also used raw instrument measurements, i.e., surface observations of precipitation, wind, and temperature. Their advantages are: 1) they verify the sensible weather parameters that we are most interested in; and 2) they verify mesoscale information. Their disadvantages are: 1) the observations must be fit to the model grid or vice versa; 2) instrument error is often a concern; and 3) it samples subgrid scale frequencies that the model can not produce. (I.e., closely spaced observations compared to the lower resolution model data create an overestimation of the error.)

## **2. Bias Correction**

Richardson (2001a) showed that, for medium-range ensemble forecasting (MREF), correcting for bias improves skill. This effect may be even greater for SREF since, as previously discussed, model deficiencies (including model bias) contribute a larger portion to the total forecast error in the short-range, before error growth from IC errors becomes very large. It is therefore critical to correct for model bias in order to realize the full potential skill of a SREF. Additionally, we found that it is difficult to analyze the results of SREF output without bias correction. In fact, our conclusion of the importance of accounting for model uncertainty in a SREF became much stronger using bias-corrected results.

In designing a bias removal method, our goal was not to pursue completely unbiased forecasts with some complex routine (e.g., multiple regression as used by Model Output Statistics) but simply to remove the bulk of the bias with an effective method and then to study the effects on ensemble performance. Scatter plots (Figure 24) of forecasts vs. observations reveal that the bias is predominantly linear and easily identifiable at a given model grid point. A

fairly simple method of using a mean bias correction is therefore appropriate. We also found that the bias is highly dependent on location, forecast lead time, flow regime, and ensemble member (i.e., model), which somewhat complicated our simple method.

Using the complete dataset, Figure 24 shows how the *MSLP*, 48-h forecast bias (defined as forecast/observation) varies for different models at the same grid point. The ngps and gasp forecasts have completely opposite biases while the ukmo forecasts (at this grid point) are nearly unbiased. Notice however that when the ukmo ICs and LBCs are used in MM5 for ACME<sup>core</sup>, the forecasts then exhibit bias (of MM5). Even though all members of ACME<sup>core</sup> use the same model, they still have different biases. In a mesoscale model, there is evidently a component of bias from both model and from the ICs and LBCs.

Figure 25 shows how *MSLP* forecast bias varies over space and lead time for a given member. One glaring fact is that the bias behaves very differently over land and ocean. The high bias over the ocean (especially the northern Pacific) is likely due to underforecasting the intensity of cyclones. Over land, there is a predominantly low bias, which could be due to incorrect heating in the boundary layer and/or problems with the reduction of pressure to sea level over high terrain. It is very evident that the biases are significant and highly dependent upon location and forecast lead time.

For a given parameter, we defined bias by

$$b_{i,j,t} = \frac{1}{N} \sum_{n=1}^N \left( \frac{f_{i,j,t}}{o_{i,j}} \right)_n \quad (26)$$

where  $N$  is the number of forecast cases in the training data,  $f_{i,j,t}$  is the forecast at grid point  $i, j$  and lead time  $t$ , and  $o_{i,j}$  is the verifying observation. This bias was then applied to a new forecast (not in the training data) to create a corrected forecast by

$$f_{i,j,t}^* = \frac{f_{i,j,t}}{b_{i,j,t}} \quad (27)$$

In general a large amount of training data (i.e., long training period) is desirable to insure a sound statistical sampling of the bias. However, we found that a long training period (e.g., the last 60 forecast cases) produced rather small improvements. This is likely due to a slow but steady shifting of the bias due to changes in the flow regime. For example, a model that typically underforecasts  $T_2$  at some location may do so with varying severity depending upon season or the prevailing synoptic situation, as evidenced in Figure 26. At the other extreme, we found that using an very short training period (e.g., the last 5 forecast cases) produced highly variable results with a mix of spectacular improvements and large degradations. This likely reflected regime shifts in which similar errors occur for several days in a row. Since it is beyond our ability to predict such shifts, we compromised on a 2-week training period to smooth out the variability.

Since we wanted to demonstrate a method that could be applied in real time, we used a running bias removal where a unique bias correction was computed each 48-h forecast period, based on model performance over the previous 2 weeks. An example training period for the forecast initialized at 00Z on 29 Jan 2003 is shown in Figure 10. Where there are missing case days, the training period is extended to always include 14 forecast/observation data pairs. The bias-corrected dataset, a subset of the full dataset, begins on 25 Nov 2002 and consists of 112 total forecast cases (Figure 10).

This bias removal technique worked quite well for *MSLP* and  $T_2$ , but not very well for  $WS_{10}$  for two reasons. One problem is that while a multiplicative bias is appropriate since  $WS_{10}$  bias appears to increase with wind speed (Figure 27a), unrealistic bias values can result for very small  $WS_{10}$  values. Secondly, unlike *MSLP* and  $T_2$ , the variance of  $WS_{10}$  errors increases with wind

speed and also becomes very large toward the 48-h lead time. Of most concern is that the errors can vary widely from case to case, often producing an inappropriate bias.

Figure 27b shows that applying the above bias removal technique can result in a severe overcorrection. This effect is greatly relieved by simply removing any instance of forecast or observed  $WS_{10}$  below 1.0 m/s, thus avoiding unrealistic bias values (Figure 27c). However, there are still an unacceptably large number of notable underforecasts, which is a larger concern for operational forecasting than overforecasting. Raising the cutoff further reduces this problem, but a problem of undersampling then arises. With a higher cutoff, the 14-day training period often contains only a few samples, making it very unreliable. We therefore chose to keep the cutoff at 1.0 m/s and reduce the resulting multiplicative bias from Equation (26) by 50%. In other words, once the bias is identified, the forecast is given an adjustment in the right direction but lessened to avoid the problem of overcorrection. We expect a slight overforecast bias to remain, but that is a better option than having a large number of underforecasts. It is the large variability of the  $WS_{10}$  errors that makes the reduced bias-correction necessary.

Figure 28 through Figure 34 show the results of our bias-correction method for all SREF systems and forecast parameters of interest. The *RMSE* and bias (forecast – analysis) were averaged over all grid points of the bias-corrected dataset. The results for the ensemble mean forecast are included for each SREF system since removing its bias is what we are really trying to do in this process. The goal is to produce more highly skilled *FP* by forcing the forecast PDF to be centered about the verification in the long-term average.

As one might expect, a larger improvement was realized where there was a larger bias. This is most evident in the *MSLP* results where the PME members are on the low extreme with small average biases and percent improvements, and the ACME<sup>core+</sup> members are on the other extreme. The lower bias of the PME members is likely due to the lower resolution and better tuning of

these large-scale models. A mesoscale model may produce more bias as it attempts to represent smaller scale phenomena with smaller grid spacing and more complex parameterizations.

The reason that ACME<sup>core+</sup> generally has larger biases and *RMSE* values compared to that of ACME<sup>core</sup> is because many of the model options selected are inferior to the standard MM5 version (Table 4). Notice that this inferiority is mostly reflected in biased error since the differences between parallel members of the two systems before bias correction are dramatic but negligible after bias correction. We concluded, by comparing the results of ACME<sup>core</sup> avn and ACME<sup>core+</sup> plus01, that this effect is due to the model option variations and not from bias introduced through our SBP perturbations. These members have nearly identical model options and the same average bias and *RMSE*, but they produce quite different solutions due to plus01's perturbed SBPs. This likely means that the SBP perturbations are performing precisely as desired, producing an equally likely solution by perturbing within uncertainty.

Notice that a model can have a shifting bias so that it displays little bias on average. Consider ACME<sup>core</sup> avn *MSLP* (Figure 29) at 24 h, which shows negligible average bias before and after correction, but a 14% improvement in *RMSE*. The explanation is that the forecasts contained opposing biases that mostly averaged out over space and time but were corrected for by our method.

There are several conclusions to be made by comparing *MSLP* bias and *RMSE* between the PME and ACME<sup>core</sup> (Figure 28 and Figure 29). The MM5 forecast from the same ICs are generally worse than the parallel large-scale model, especially for the superior models such as avn and ukmo. This may be partly due to the effect of MM5's higher model resolution artificially increasing *RMSE*, but it is more likely due to the fact that the global models can more accurately predict the development of large-scale weather systems. Such information is only weakly translated into the MM5 solution through LBC updates so synoptic waves within the MM5



domain can drift off considerably from the large-scale model's solution. Also note the similarity in bias and relative *RMSE* of the parallel component members of the PME, ACME<sup>core</sup>, and ACME<sup>core+</sup> (Figure 28 – Figure 30). This likely indicates that, for a predominantly synoptic-scale parameter such as *MSLP*, the primary source of forecast error is the ICs since applying the same ICs to different models makes only small differences in the error.

For  $T_2$  bias and *RMSE* of ACME<sup>core</sup> (Figure 31) there is almost no difference among the various members. This indicates that for  $T_2$ , a primarily mesoscale parameter, forecast error is mostly influenced by the model and not the ICs since the error is virtually the same no matter what IC is applied. This conclusion is reinforced in Figure 32 where the different models of ACME<sup>core+</sup> do exhibit notable variations in bias and *RMSE*. Lastly, the fact that there is very little growth in the error with forecast lead time is a third indication of the predominance of model error. (This will be discussed further in the next chapter.)

An interesting result of the  $T_2$  bias correction is the disparity between the 12/36-h bias and the 24/48-h bias. Since all forecasts were made at 00Z (5 PM local time), the difference is for the late night bias vs the late afternoon bias. Evidently, MM5 greatly underforecasts the late afternoon temperature from the daytime heating. This is true for the standard MM5 version and even more pronounced for some members of ACME<sup>core+</sup>. The late night  $T_2$  bias is much weaker and varies depending on MM5 version. The strong late afternoon bias points to a serious deficiency in the radiation and PBL schemes.

The results of the  $WS_{10}$  bias correction (Figure 33 and Figure 34) are unimpressive compared to those of *MSLP* and  $T_2$ . The higher variability in  $WS_{10}$  errors makes any bias removal scheme less effective. Furthermore, the effect of the 50% reduction in the multiplicative bias is also evident as much of the overforecast bias remains after correction. All of our attempts to fully remove the bias resulted in degradations of *RMSE* (not shown).

The variability of the  $WS_{10}$  bias and  $RMSE$  among the members is larger than with  $T_2$  but not as large as with  $MSLP$ . This suggests that for  $WS_{10}$ , the source of the error is a fairly even mix of model and IC. This makes sense since surface winds are determined by the large-scale pressure gradient at the surface (which is mostly determined by the ICs) and by mesoscale features such as local terrain and heating (which are determined by model physics). Examining the source (model vs. IC) is an important issue in this research and will be explored further in the next chapter.

Lastly, to confirm the value of this bias removal technique, we used observation-based verification over a one week period to evaluate both uncorrected and bias-corrected forecasts. This is an indirect way to determine the quality of the gridded analysis used in the bias correction. It is possible that our bias correction simply adjusted the forecasts toward a poor or biased representation of truth. If the grid-based bias removal also improves the forecast with respect to station observations, we can be more confident in the quality of the gridded analysis. The big advantage of a grid-based vs. an observation-based bias removal is that the grid-based provides a domain-wide improved forecast, rather than only at the limited areas covered by observations.

Figure 35 shows that for  $MSLP$  the grid-based bias removal does work rather well with respect to station observations. Although the negation of bias and percent improvement are not as impressive as in Figure 29, they are still quite positive. This result leads us to conclude that the centroid analysis is in good agreement with station observations. One significant disparity between Figure 29 and Figure 35 is the much larger  $RMSE$  for the observation-based verification. This is most likely due to the concentration of station observations over land for Figure 35 (where  $MSLP$  is more variable), whereas Figure 29 was made using the entire 36-km domain.

The observation-based verification results for  $T_2$  are mixed. Figure 36 shows that while we obtained excellent results for negating the bias at all lead times, we were only able to improve  $RMSE$  in the late afternoon times and actually degraded the forecast in the late night lead times.

The reason for the nighttime degradation is likely due to the higher variability of  $T_2$  errors, making it possible to correct a bias but difficult to improve *RMSE*. Additionally, the RUC20 analysis probably does not agree well with the station observations at night since the RUC20 model may have serious deficiencies in modeling the nighttime boundary layer.

### 3. Forecast Probability Calculation

Back in Chapter I.A, we introduced the idea that potentially the most valuable application of EF is the production of *FP* of some forecast event. This is because it combines all the EF information into a single product, encapsulating the forecast uncertainty and providing a product useful in decision making. In this section we will describe the *FP* calculation method that we employed. To simplify the discussion somewhat, the equations and sample calculations will all be for the probability of the verification exceeding the event threshold.

There are many possible ways to calculate *FP* from an EF. In Chapter I.A we described how one could use the appropriate area under a PDF that was directly fitted from the ensemble. This method, revisited in Figure 37 for a hypothetical forecast PDF of  $WS_{10}$ , would only be effective for a very large ensemble. For an ideal ensemble of infinite size, the resulting  $FP = 77.1\%$  for an event threshold of 20.0 kt represents the genuine probability of occurrence. Note that if the ensemble is nonideal but infinite in size, this method does not guarantee skillful *FP*. To achieve high resolution (i.e., sharp forecasts) and high reliability ( $FP \approx ORF$ ), the EF system still has to meet the other demands of properly accounting for analysis and model uncertainty.

For practical purposes, a different method is required to obtain *FP* since it is not normally possible to reliably fit a PDF to an ensemble of finite size and to a distribution of unknown shape. Consider a simulated ensemble of  $WS_{10}$  forecasts at some grid point, created by drawing eight random, ordered samples from the true forecast PDF (thick curve in Figure 37):

$$WS_{10} = \{16.5, 21.1, 23.3, 25.3, 27.4, 34.4, 40.2, 47.8 \text{ kt} \}$$

For demonstration purposes, we fit a continuous PDF (thin curve in Figure 37) to this data to show how low sampling alone creates an erred  $FP = 81.6\%$ , but again, such a technique is not practical because of the difficulties in making a reliable fit.

The most common method (used in current, operational EF) to calculate  $FP$  is often called *democratic voting* (DV). As the name implies, each ensemble member gets an equal vote on what the true state of the atmosphere may be. Mathematically, the probability of the verification ( $V$ ) occurring above the event threshold ( $\tau$ ) is simply found by

$$P(V > \tau) = \frac{1}{n} \sum_{i=1}^n (1 \text{ if } x_i > \tau, 0 \text{ if } x_i \leq \tau) \quad (28)$$

where  $x_i$  is the value of the  $i^{\text{th}}$  ensemble member. Using the same  $\tau$  as above,  $FP = 7/8 = 87.5\%$  since seven of the forecasts were greater than 20.0 kt. The large error of 10.4% compared to the genuine  $FP$  of 77.1% is partly due to the small sampling but is also a result of a systematic problem with DV.

DV effectively bins  $FP$  into  $n+1$  possible values (the topmost values in Figure 38). There is nothing necessarily wrong with binning  $FP$ , but for DV the resulting bin values are fixed in a biased way with respect to the ordered EF values. The gaps (ranges of values between two members) among the ordered EF members should be considered to be an evenly spread continuum of probability since on average the members represent evenly divided quantiles. The horizontal arrows in Figure 38 show how possible positions of an event threshold get binned. DV effectively pushes  $FP$  toward the extreme values, so that high  $FP$  is normally overforecast and low  $FP$  is normally underforecast. This exacerbates the problem of low sampling as we will demonstrate below.

An alternative method which we adapted from Hamill and Colucci (1997) is called uniform ranks (UR). In the lower part of Figure 38, UR begins by uniformly breaking up the total probability into  $n+1$  ranks that match the possible rank positions of the event threshold when pooled with the ordered EF members. Similar to DV, the probability from the ranks that exceed the event threshold is summed. Then, rather than simply add on half the probability from the rank where the thresholds occurs, we add on a fraction of the probability proportional to the distance from  $\tau$  to the surrounding members' values. For a  $\tau$  with a ranking of  $i$  (when  $1 < i \leq n$ ) among the ensemble members, the probability of the verification occurring between  $\tau$  and the  $i^{\text{th}}$  member is:

$$P(\tau < V < x_i) = \left( \frac{x_i - \tau}{x_i - x_{i-1}} \right) \frac{1}{n+1} \quad (29)$$

This procedure assumes that the random variable is uniformly distributed between ensemble members. In our example, the result is  $FP = 7/9 + [(21.1 - 20.0) / (21.1 - 16.5)] * 1/9 = 80.4\%$ , a value much closer to the genuine  $FP$  of 77.1% compared to the DV  $FP$  of 87.5%. Such an improvement is not consistently the case for the two methods, but UR is a superior method on the whole.

The biggest improvement of UR over DV is for extreme  $FP$  values, i.e., when  $\tau$  is ranked 1 or  $n+1$  when pooled with the ensemble members. Continuing with the same example but now with a  $\tau = 50.0$  kt, DV would give  $FP = 0.0\%$  since all the forecasts are below the threshold. However, since the discrete members of the EF are actually representing a PDF and  $\tau$  is so close the largest member, there is still a nonnegligible chance that the verification will exceed  $\tau$ .

In UR, we calculate the fraction of probability from the outside ranks with a separate procedure (Figure 39a). As with the interior ranks, the probability is found by taking a portion of probability in the outside rank based on the numerical distance between the highest member and

$\tau$ . However, using a linear proportion is inappropriate since we are dealing with the tails of the PDF. Additionally, there is no  $(n+1)^{\text{th}}$  EF member with which to calculate a linear proportion. Therefore, the total probability of rank  $n+1$  is considered to be the upper extreme end of the sample's theoretical Gumbel cumulative density function (CDF) (Wilks, 1995):

$$F(x) = \exp\left(-\exp\left(\frac{\xi - x}{\beta}\right)\right) \quad (30)$$

$$\hat{\beta} = \frac{s\sqrt{6}}{\pi} \quad \hat{\xi} = \bar{x} - \gamma\hat{\beta}$$

where  $x$  is the random variable, and  $\beta$  and  $\xi$  are the Gumbel parameters estimated using the sample standard deviation  $s$  and the sample mean  $\bar{x}$ . The Gumbel distribution was used because of its ability to characterize extreme events (Hamill and Colucci, 1997; Wilks, 1995). The probability of the verification occurring above the event threshold is then:

$$P(V > \tau) = \left( \frac{1 - F(\tau)}{1 - F(x_n)} \right) \frac{1}{n+1} \quad (31)$$

where  $F(\tau)$  is the Gumbel CDF value at  $\tau$ ,  $F(x_n)$  is the Gumbel CDF value at the value of the highest ranked ensemble member,  $x_n$ . After fitting our example EF to the Gumbel, we calculate  $FP = [(1 - F(50.0)) / (1 - F(47.8))] (1/9) = 8.5\%$ . This is a low but significant chance of occurrence for which DV would assign an  $FP$  of 0.0%.

The opposite extreme of  $\tau$  occurring in rank 1 (i.e.,  $\tau$  falls below the lowest ensemble member) is handled in a similar fashion by reversing the Gumbel CDF since it is the right tail that represents extreme events so well. For random variables such as  $WS_{10}$  that are bound by 0.0 on the left, we mimic a fixed CDF with an exponential, thus assuring that probability drops to zero as required:

$$P(V > \tau) = \left(1 - \left(\frac{\tau}{x_1}\right)^3\right) \frac{1}{n+1} \quad (32)$$

where  $x_1$  is the value of the lowest ranked ensemble member.

When the ensemble produces a poor estimate of the forecast PDF, UR can not provide a much better approximate *FP*. In such cases, which of course occur more frequently for smaller  $n$ , the *FP* from both UR and DV suffers equally. (The only way around that problem is to increase ensemble size since a poor sampling can not be identified a priori.) Ironically though, UR is a more dramatic improvement over DV when  $n$  is small since DV suffers more for smaller  $n$ . Therefore, the real improvement of UR over DV is for small  $n$  on cases that are reasonable approximations to the forecast PDF. On average, over a large number of realizations, UR produces superior *FP* since it either performs the same or better than DV.

Richardson (2001b) showed that the result of undersampling on *FP* is an overconfident EF. Since the tails of the PDF are less likely to be represented, high *FP* values are normally overforecast and low *FP* values are normally underforecast. This effect is revealed in a reliability diagram by a curve with a clockwise tilt. What Richardson (2001b) failed to realize is that DV, which he used to calculate *FP*, exacerbates the problem with its biasing of *FP* toward the extreme values.

To demonstrate this fact, we performed a sampling experiment similar to that of Richardson (2001b) where a perfect ensemble was simulated by taking an observation and  $n$  random draws from the same  $WS_{10}$  PDF. The event threshold applied was again  $\tau = 20.0$  kt and the PDFs were similar to Figure 37 but allowed to vary from case to case to get a full range of *FP*. *FP* was calculated by both DV and UR using a set of  $10^6$  simulated forecast cases. Figure 40 shows that both methods result in an overconfident EF, but UR is a dramatic improvement over DV for small  $n$ . As  $n$  increases, both methods approach perfect reliability and the improvement by UR

diminishes. This experiment also provides us a way to estimate the expected Brier Skill score (*BSS*) improvement by the ACME system ( $n = 17$ ) over ACME<sup>core</sup> ( $n = 8$ ). Repeating the experiment with those ensemble sizes we found that there was an increase in *BSS* of  $\sim 0.03$ .

While the UR method produces a better estimate of *FP*, the result is just as uncalibrated as DV. Both methods assume that the ensemble members are all equally likely and that there are no systematic errors. Figure 39b shows how the UR method can be changed into the *weighted ranks* method to account for systematic errors (Eckel, 1998). Instead of multiplying by  $1/n+1$  in Equations (29), (31), and (38), we multiply by the historical probability of verification occurring in that rank. This is provided by a verification rank histogram, a record of where the verification has occurred among the ordered EF members over many past cases. By using rank probability based on past performance of the ensemble, systematic errors in the ensemble are compensated for. In Figure 39b, the ensemble is evidently underdispersive so there is a greater chance of verification occurrence in the last rank. The fraction of the rank's probability is calculated as in UR, but the final value of *FP* is now higher, reflecting the greater odds that the event will occur given this particular EF.

The weighted ranks method produces a more reliable and calibrated *FP* (Hamill and Colucci, 1997; Eckel and Walters, 1998). We did not, however, apply this technique because *FP* calibration is not a specific issue of this research. We chose to be satisfied with the results of the UR method.



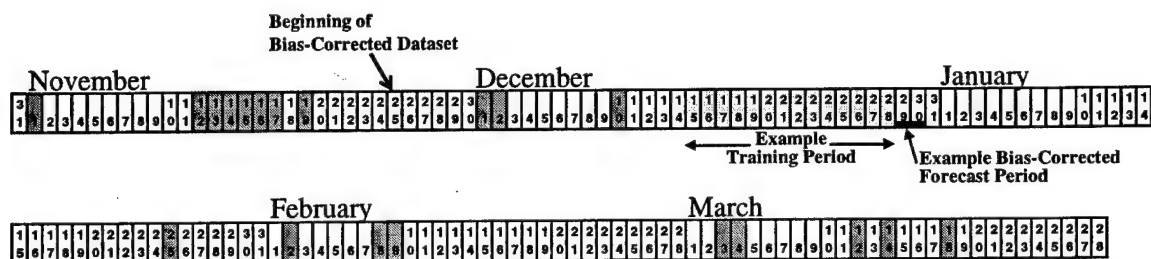


Figure 10. The 129 forecast case days of the research dataset over the 2002-2003 cool season. A 48-h forecast cycle was initialized at 00Z on each date. Darkly shaded dates contain at least one incomplete or missing member of one of the ensemble systems, so were dropped from the dataset. The lightly shaded 2-week period is an example training period that was used to compute a bias correction for the indicated example forecast period. The bias-corrected dataset consists of 112 cases, which are the complete cases beginning 25 Nov 2002.

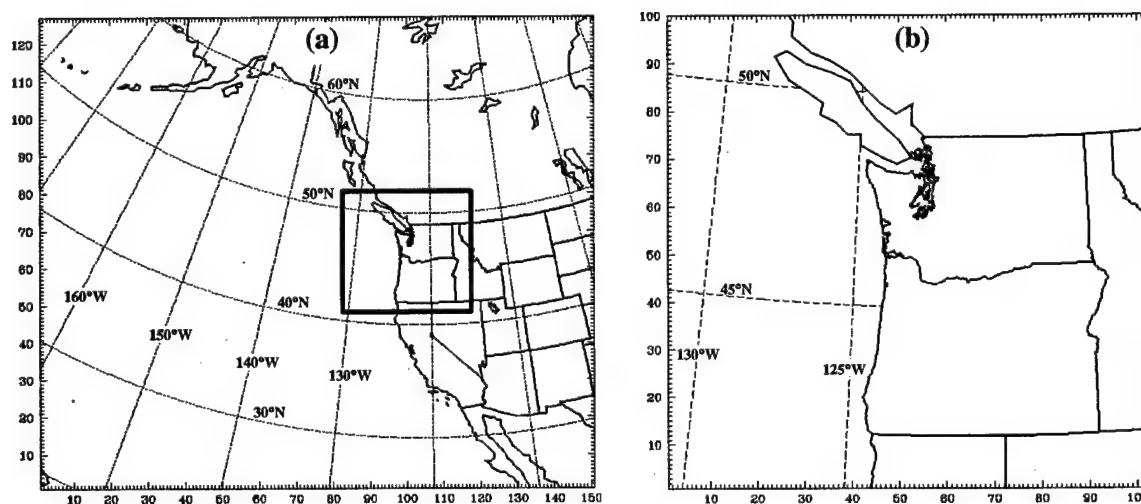


Figure 11. Grid domains (Lambert conformal projections) of the SREF systems. (a) 151x127, 36-km resolution domain. (b) 103x100, 12-km resolution domain.

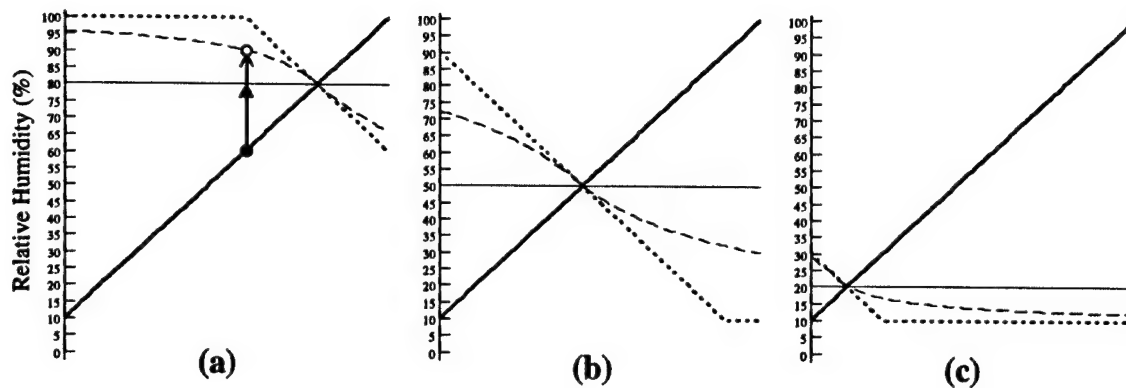


Figure 12. Display of the mirroring of  $RH$  for three different values of the centroid  $RH$  (thin solid line): (a)  $RH_C = 80\%$ , (b)  $RH_C = 50\%$ , and in (c)  $RH_C = 20\%$ . The thick solid line gives all possible values of  $RH_A$  (moisture analysis). The dotted line is the resulting mirrored  $RH$  with truncation of bad values. The dashed line is the Zeno-mirrored  $RH$  value.

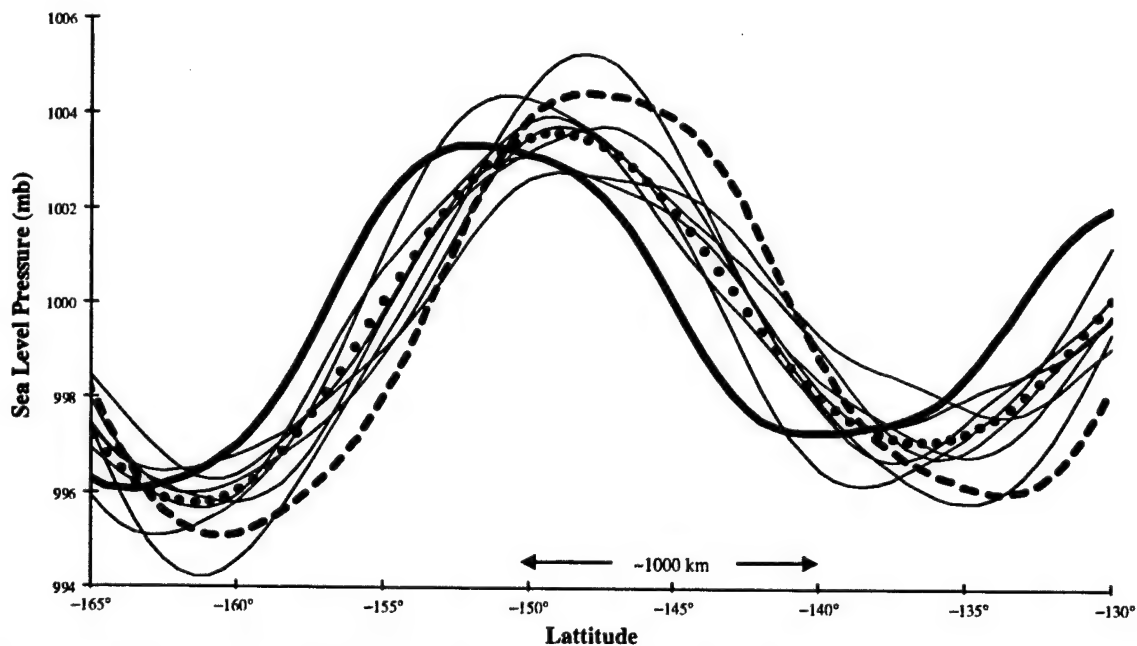


Figure 13. 2-D demonstration of the mirroring technique. The solid lines represent eight analyses of  $MSLP$  across a latitude line and the dotted line is the centroid. The thick line is the analysis used to produce the example mirrored IC (dashed line).

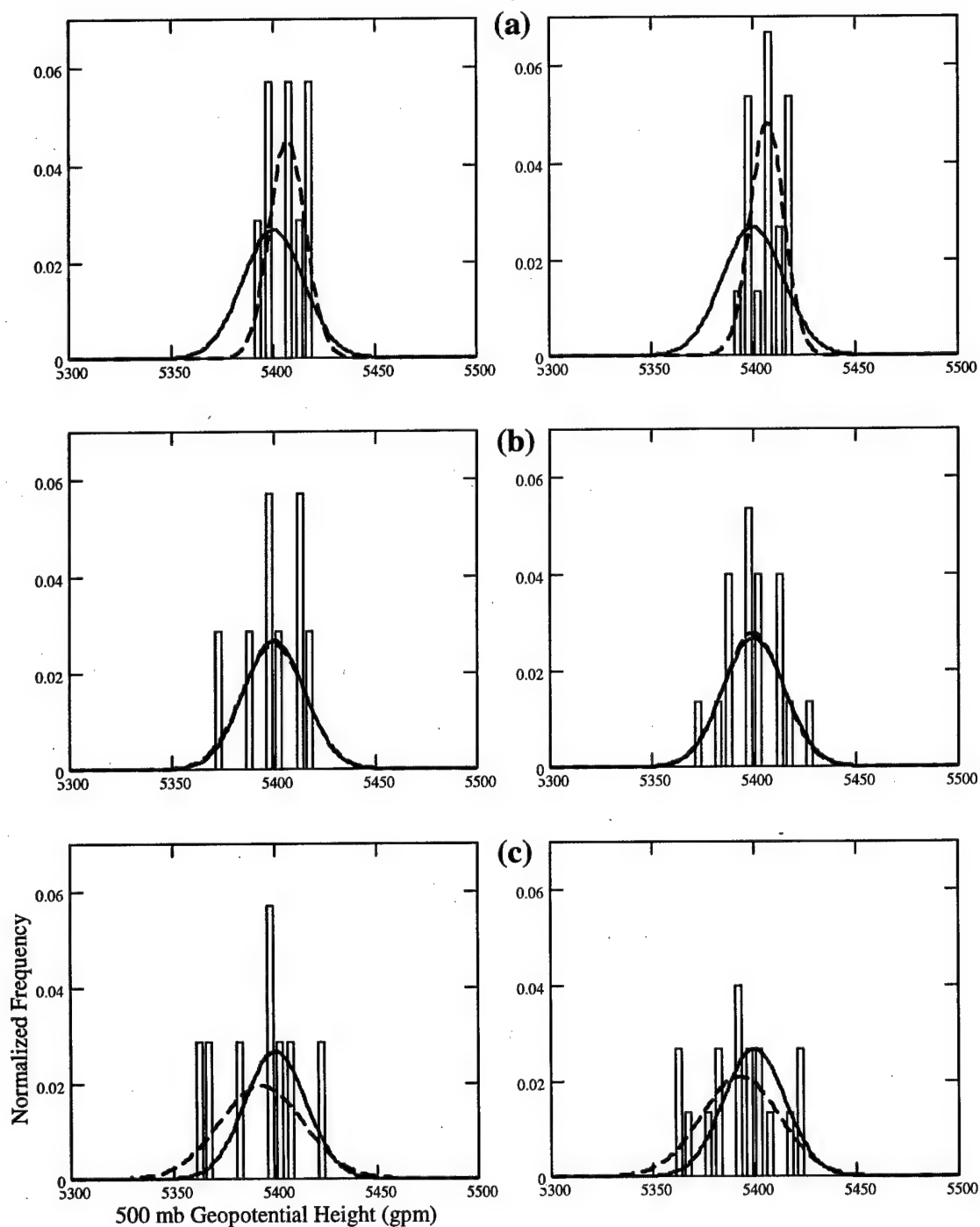


Figure 14. Simulated EF attempts to represent a hypothetical forecast PDF (solid curve). The plots on the left are normalized histogram (class interval size = 5.0 gpm) of eight random samples and their fitted normal (dashed curve), representing  $ACME^{core}$ . The plots on the right represent an expansion of the  $ACME^{core}$  plots into the full ACME (core, centroid, and the mirrors). (a) A case with  $\bar{x}$  too big and  $s$  too small. (b) A case of excellent reproduction of the PDF. (c) A case with  $\bar{x}$  too small and  $s$  too big.

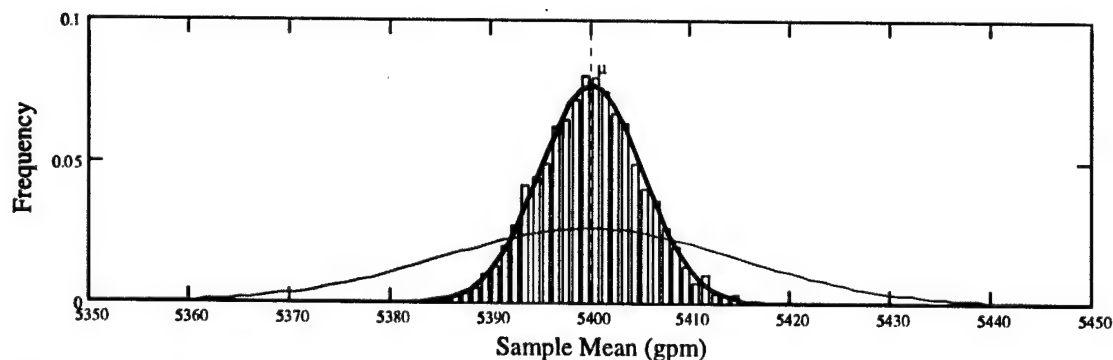


Figure 15. Sampling distribution of  $\bar{x}$  using 5000 samples of size  $n=8$ . The histogram for the sampling distribution of  $\bar{x}$  matches up well with the sampling theory PDF (thick solid curve) since we have such a large number of samples. The wider, thin curve shows the forecast PDF that the samples were drawn from ( $\mu = 5400$  gpm and  $\sigma = 15$  gpm).

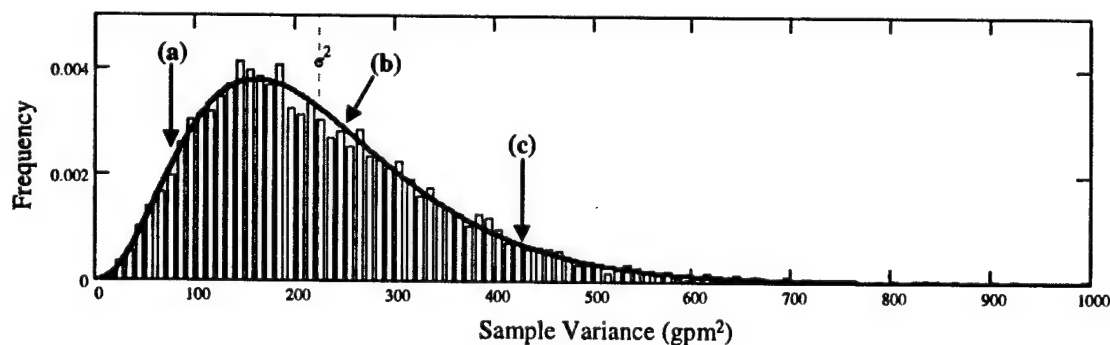


Figure 16. Sampling distribution of  $s^2$  from 5000 samples of size  $n = 8$ , similar to Figure 15 except the sampling theory PDF (thick solid curve) is a  $\chi^2$  distribution. The average  $s^2$  is  $223.7 \text{ gpm}^2$  and the standard deviation is  $130.6 \text{ gpm}$ . For reference, the variances of the three cases in Figure 14 are indicated along the distribution.

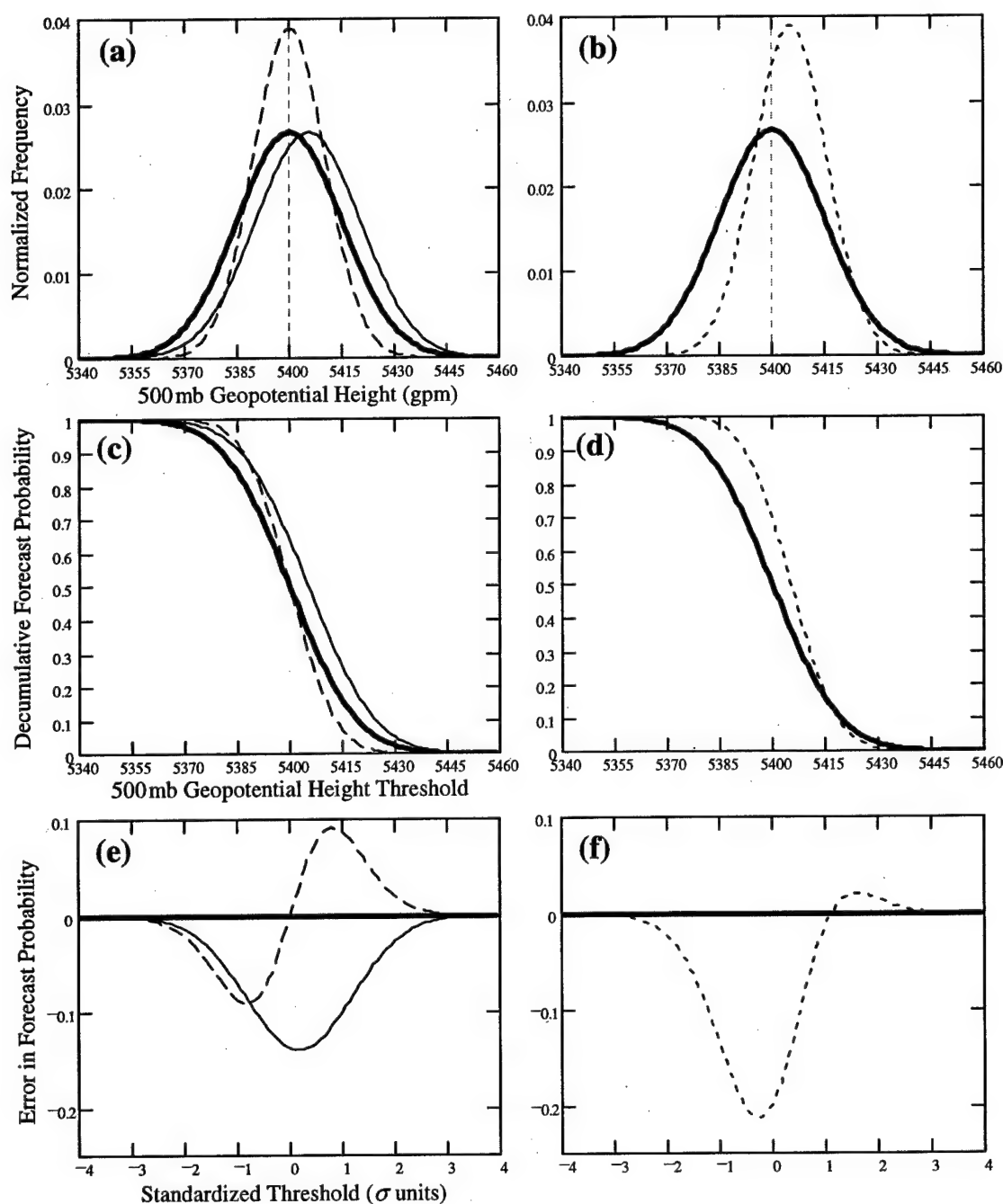


Figure 17. Impact on forecast probability of the standard error in mean and variance of an idealized eight member ensemble. The thick solid curve is for the true (i.e., correct) forecast PDF. The thin solid curve is for the standard error in the mean. The long-dashed curve is for standard error in the variance. The short-dashed curve on the right hand side panels is for the combined effect of both errors. Panels (a) and (b) are the PDFs, panels (c) and (d) are the decumulative density functions, and panels (e) and (f) resulting errors in *FP* vs. possible event threshold (standardized to  $\sigma$  units by subtracting  $\mu$  then dividing by  $\sigma$ ).

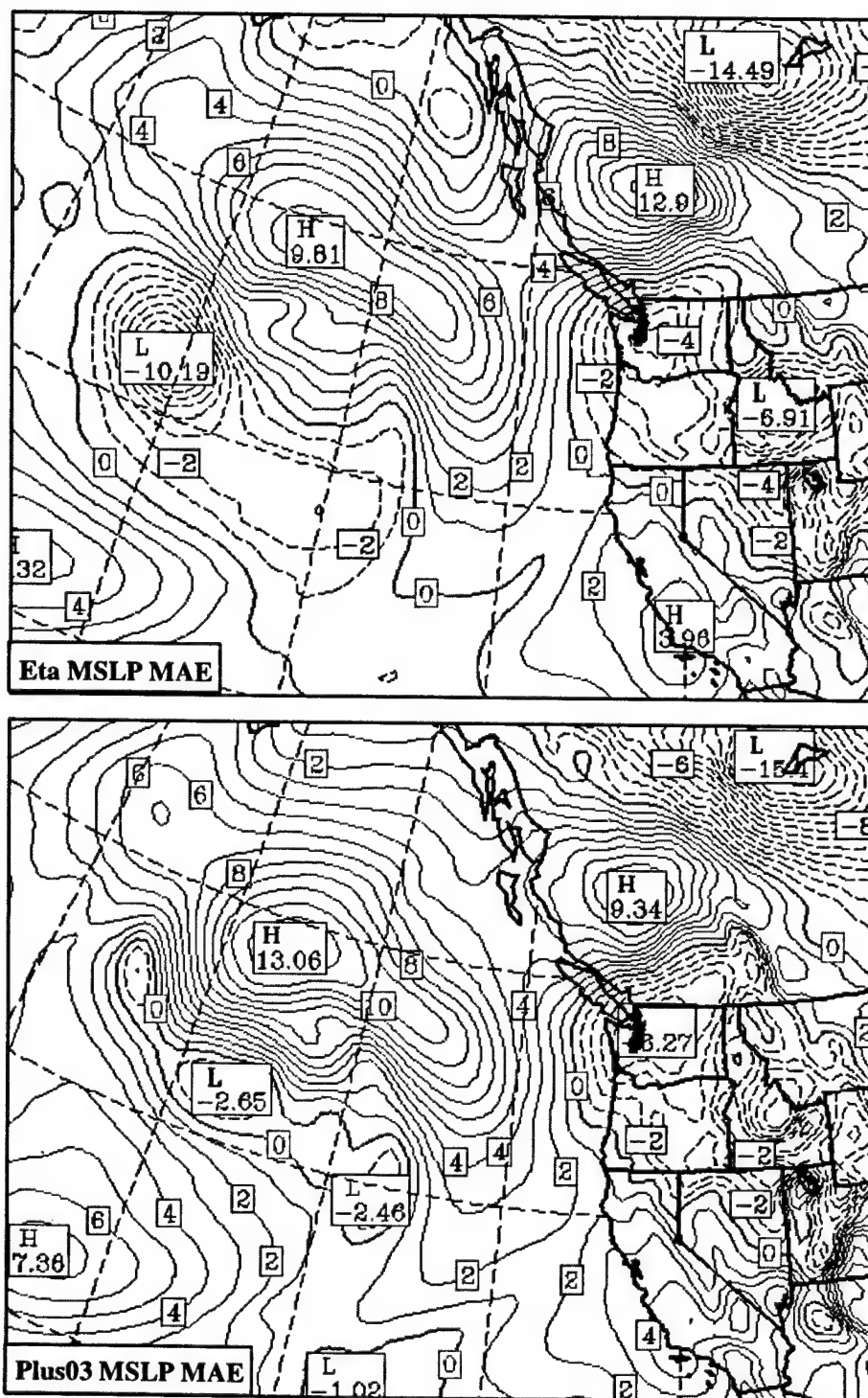
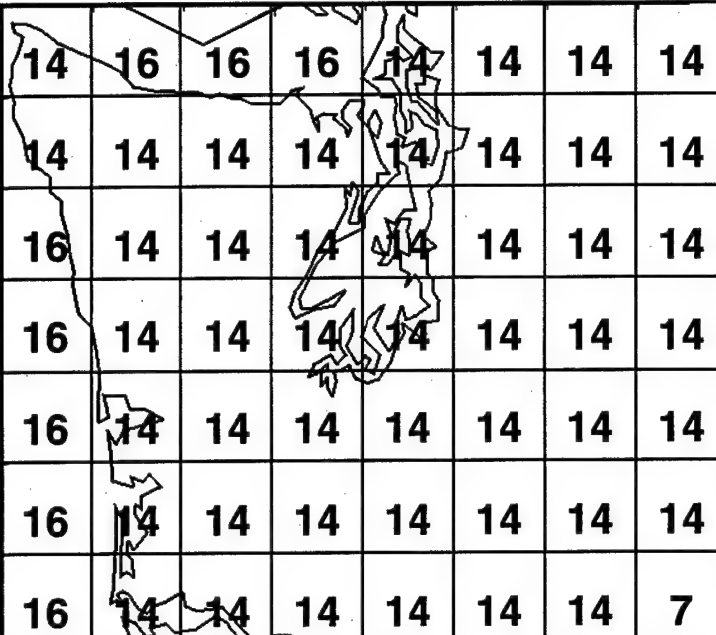


Figure 18. Mean absolute error (MAE) from a single 48-h *MSLP* forecast of the eta member of  $ACME^{core}$  and the plus03 member of  $ACME^{core+}$ . (Note that both model runs began with the same IC.) The eta solution has less error over most of the domain, but there are notable regions where plus03 performed better, such as over British Columbia.



14	16	16	16	14	14	14	14
14	14	14	14	14	14	14	14
16	14	14	14	14	14	14	14
16	14	14	14	14	14	14	14
16	14	14	14	14	14	14	14
16	14	14	14	14	14	14	14
16	14	14	14	14	14	14	7

Figure 19. Section of the 36-km resolution grid domain showing MM5 land use number over western Washington. Land use #7 is called “grassland”, #14 is called “evergreen needleleaf forest”, and #16 is called “water bodies”.

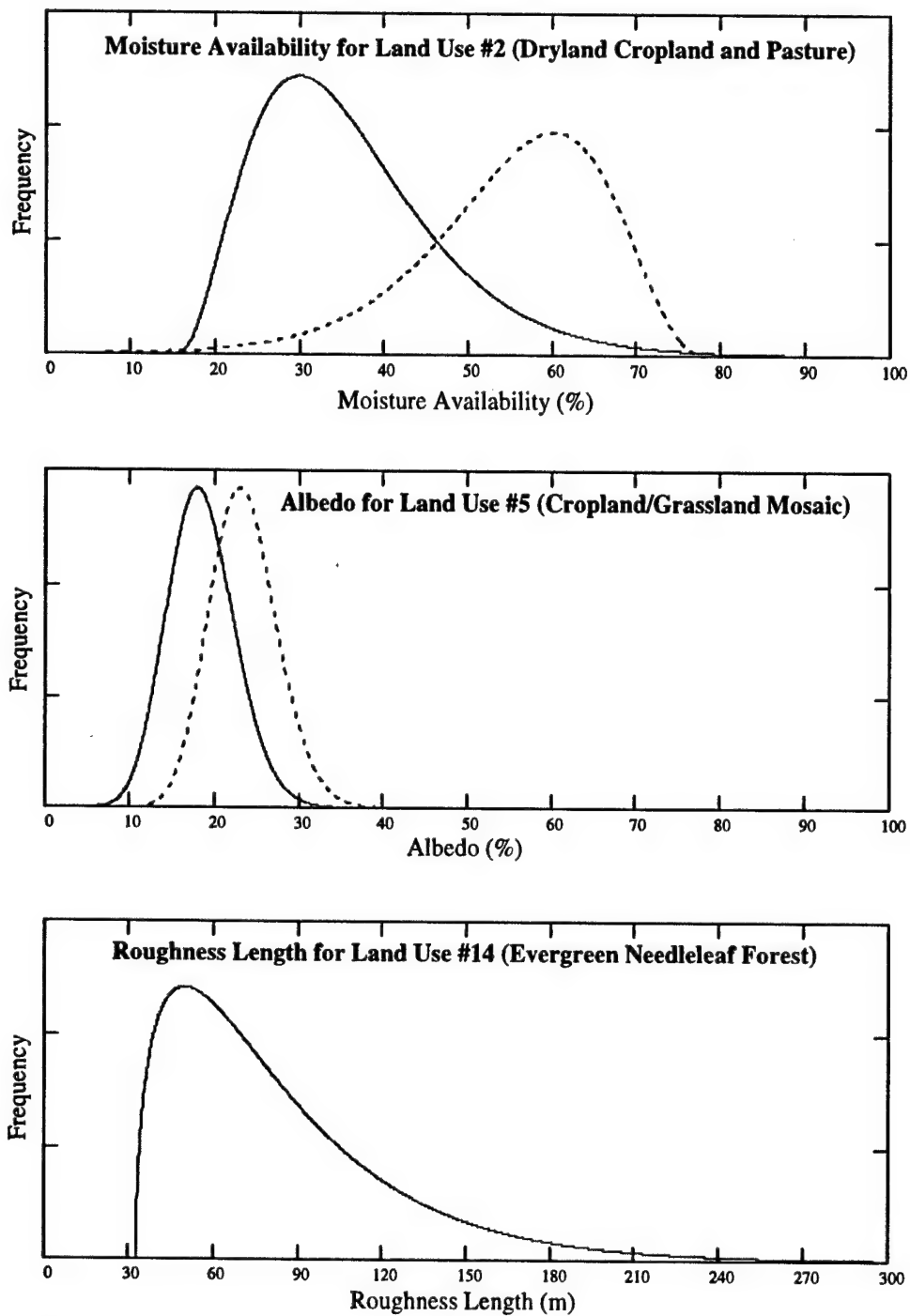


Figure 20. Sample surface boundary parameter PDFs. The solid curve is for summer and the dashed curve is for winter. The peaks of the curves correspond to the parameter values in the standard MM5 land use table. For a wider PDF there is more uncertainty in the value of the SBP.



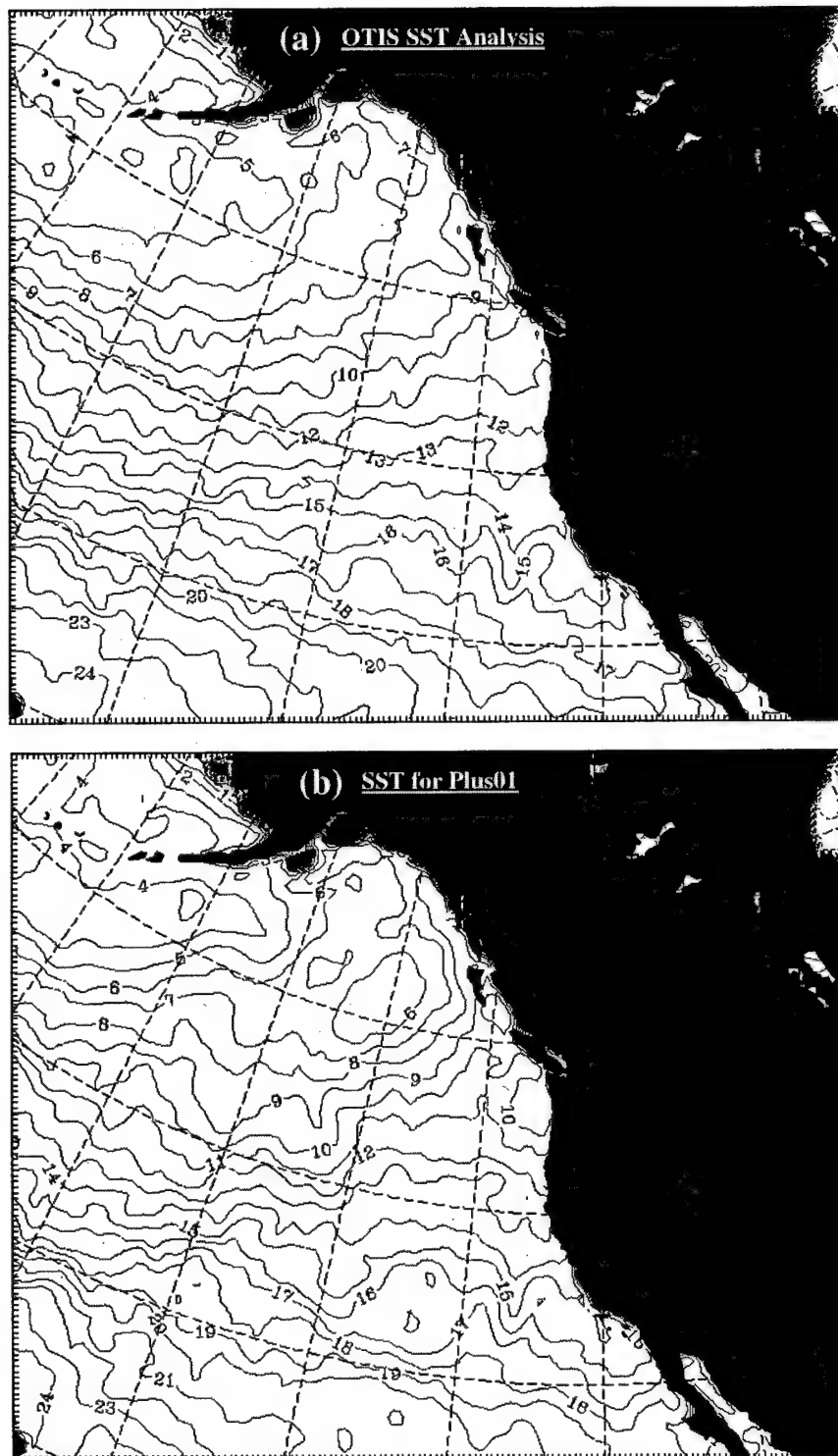


Figure 21. Example SST (in  $^{\circ}\text{C}$ ) fields from 8 Jan 2003. (a) The unperturbed field used by all ACME<sup>core</sup> members. (b) The perturbed field used by member plus01 of ACME<sup>core+</sup>, made by applying the perturbation shown in Figure 22 to (a).

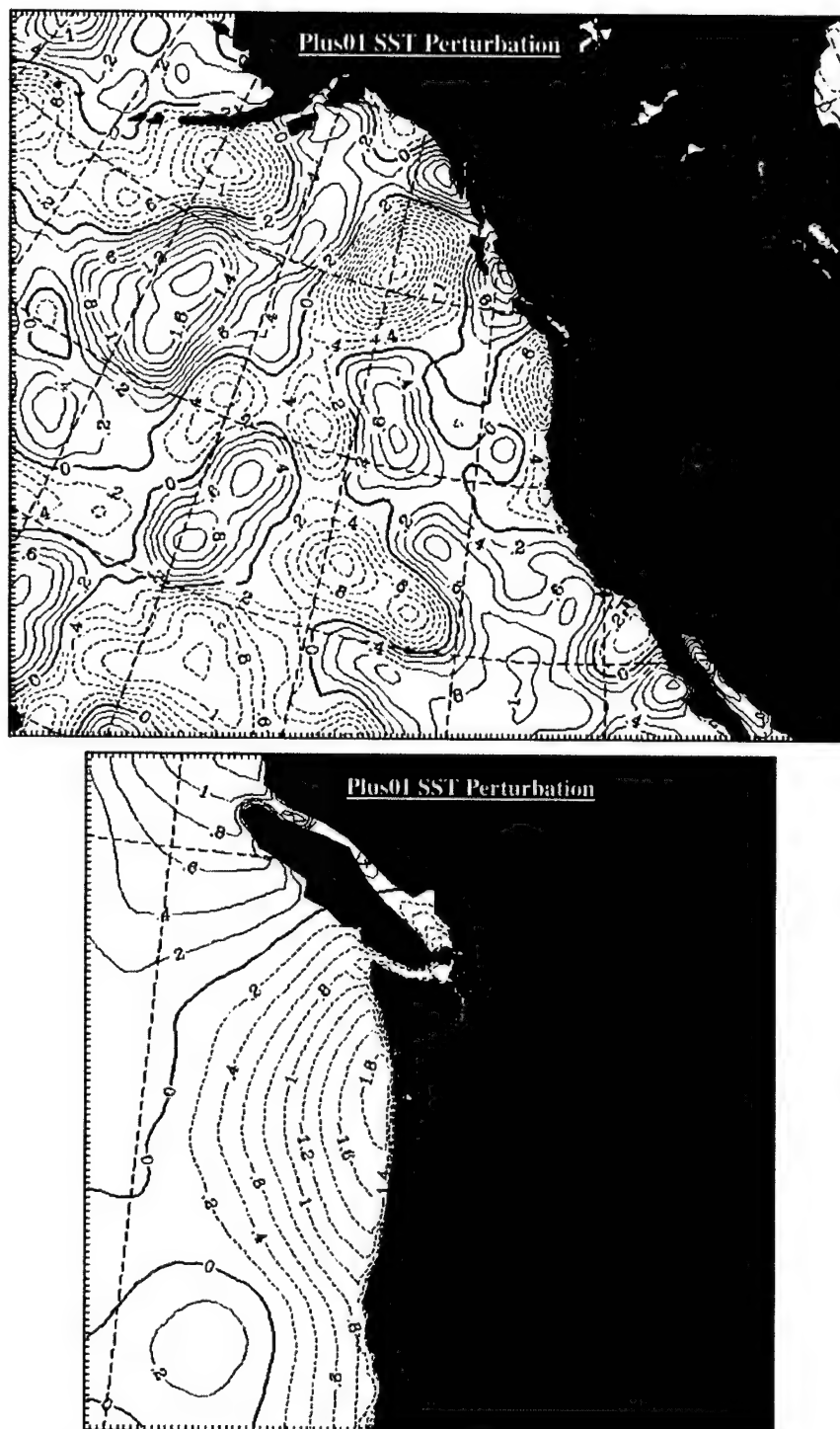


Figure 22. Plot of the SST perturbations for member plus01's outer domain and inner domain, which are made to match up. Isopleths are positive (solid) and negative (dashed) perturbation values. The apparently high gradient at the shoreline is an artifact of the plotting routine and not present in the actual perturbations.

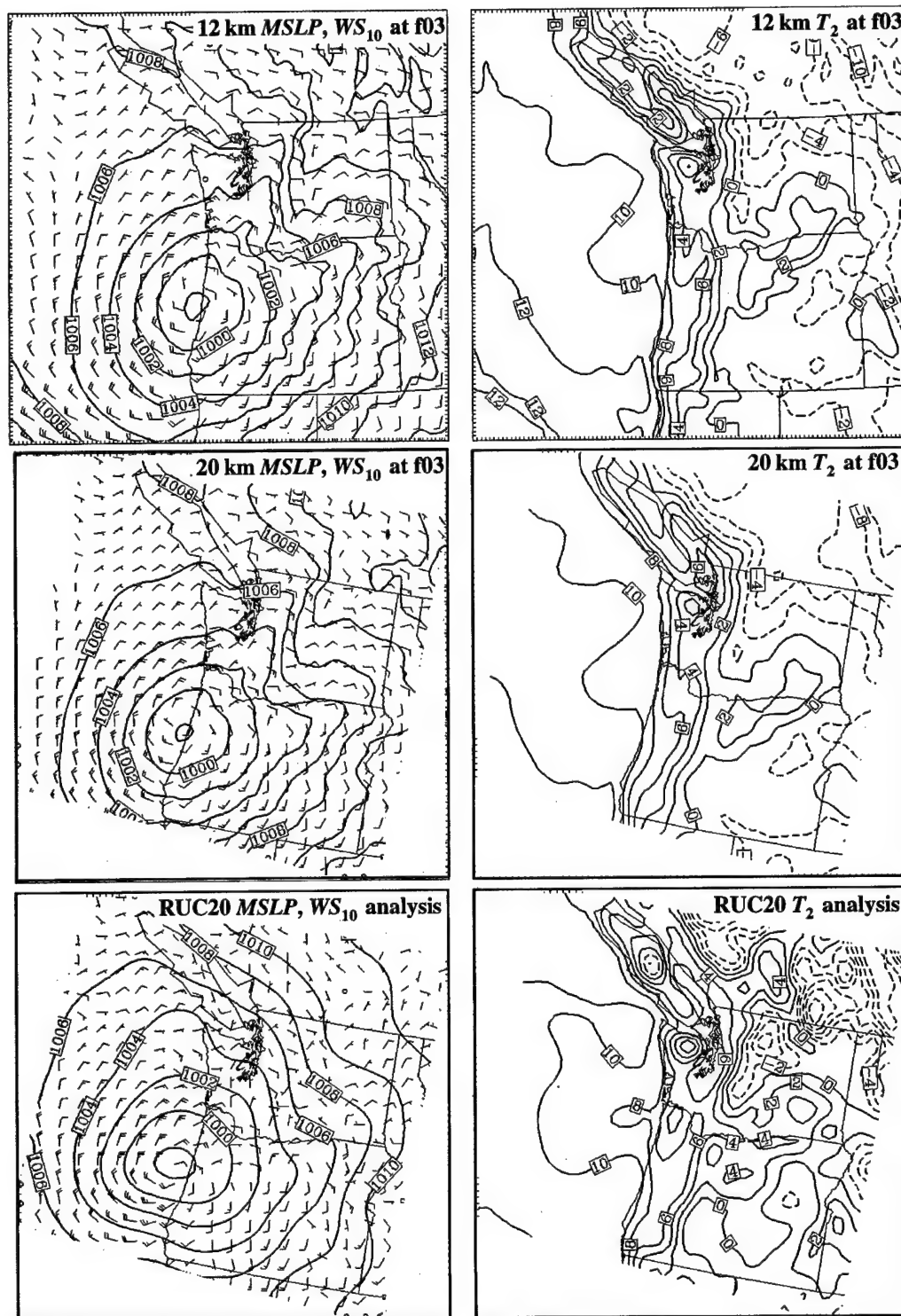


Figure 23. Sample ACME<sup>core</sup> avn 3-h forecast data taken from the 12-km MM5 grid (top plots) and fit to the 20-km RUC20 grid (middle plots), valid 3Z, 21 Dec 2002. Left column plots are  $MSLP$  and  $WS_{10}$  and right column plots are  $T_2$ . The bottom plots are the RUC20 analysis data used to verify the middle plots.

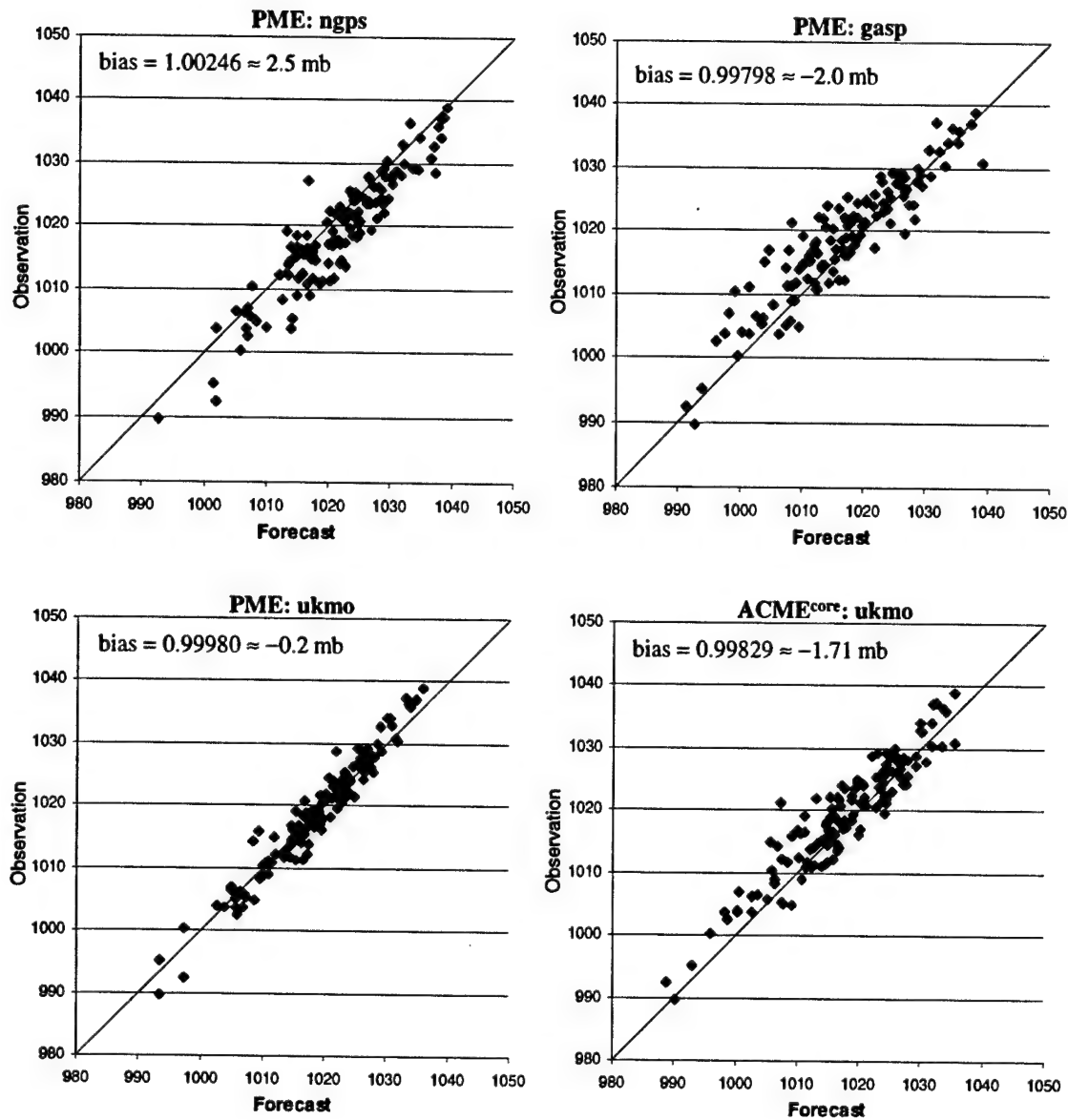


Figure 24. Scatter plots of 36-h forecast *MSLP* vs. centroid-analysis verification at point 111, 69 in the 36-km domain, a grid point in eastern Washington.

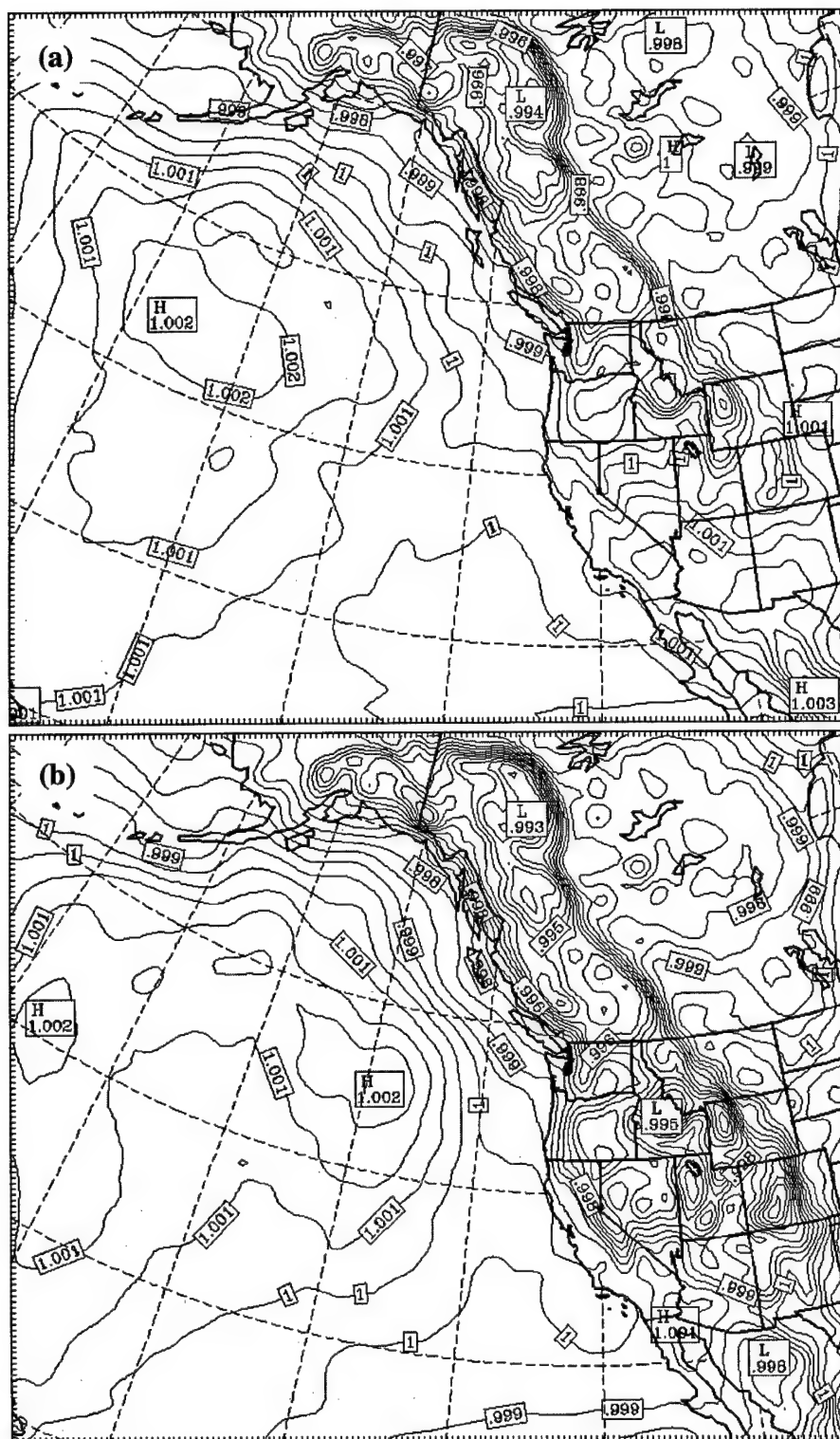


Figure 25. MSLP bias (multiplicative) for the avn member of ACME<sup>core</sup> at forecast lead time of (a) 24 h, and (b) 36 h.

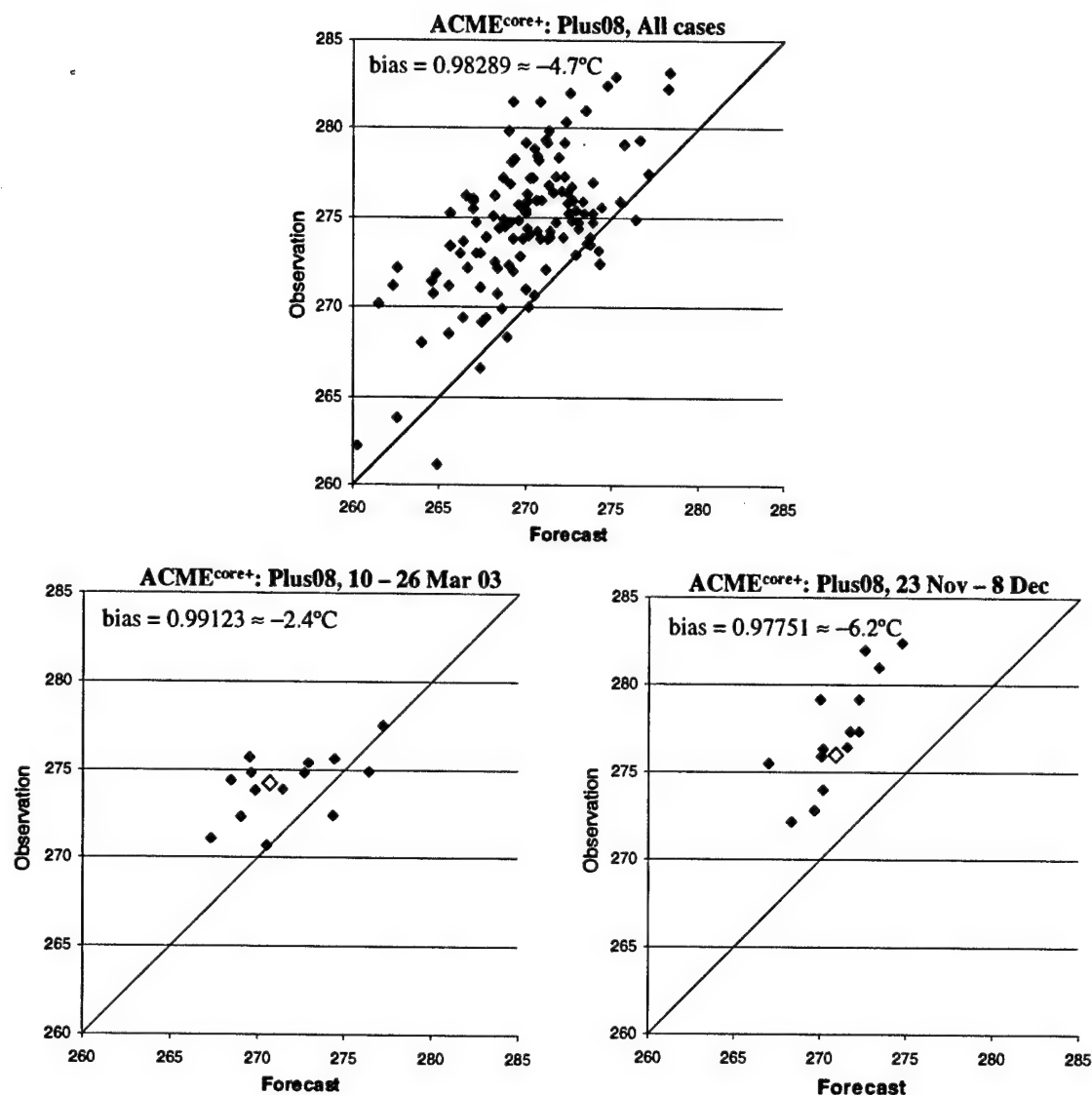


Figure 26. Scatter plots of 24-h forecast  $T_2$  vs. RUC20 verification at point 50,50 in the 20-km domain, a grid point in southern British Columbia. The top plot includes all 129 case days and the two lower plots are for subset, 14-day periods, as indicated. The open diamond in the two lower plots is the next, sequential forecast (after the 14-day period) showing that its likely bias is normally more closely related to that of the recent past cases.

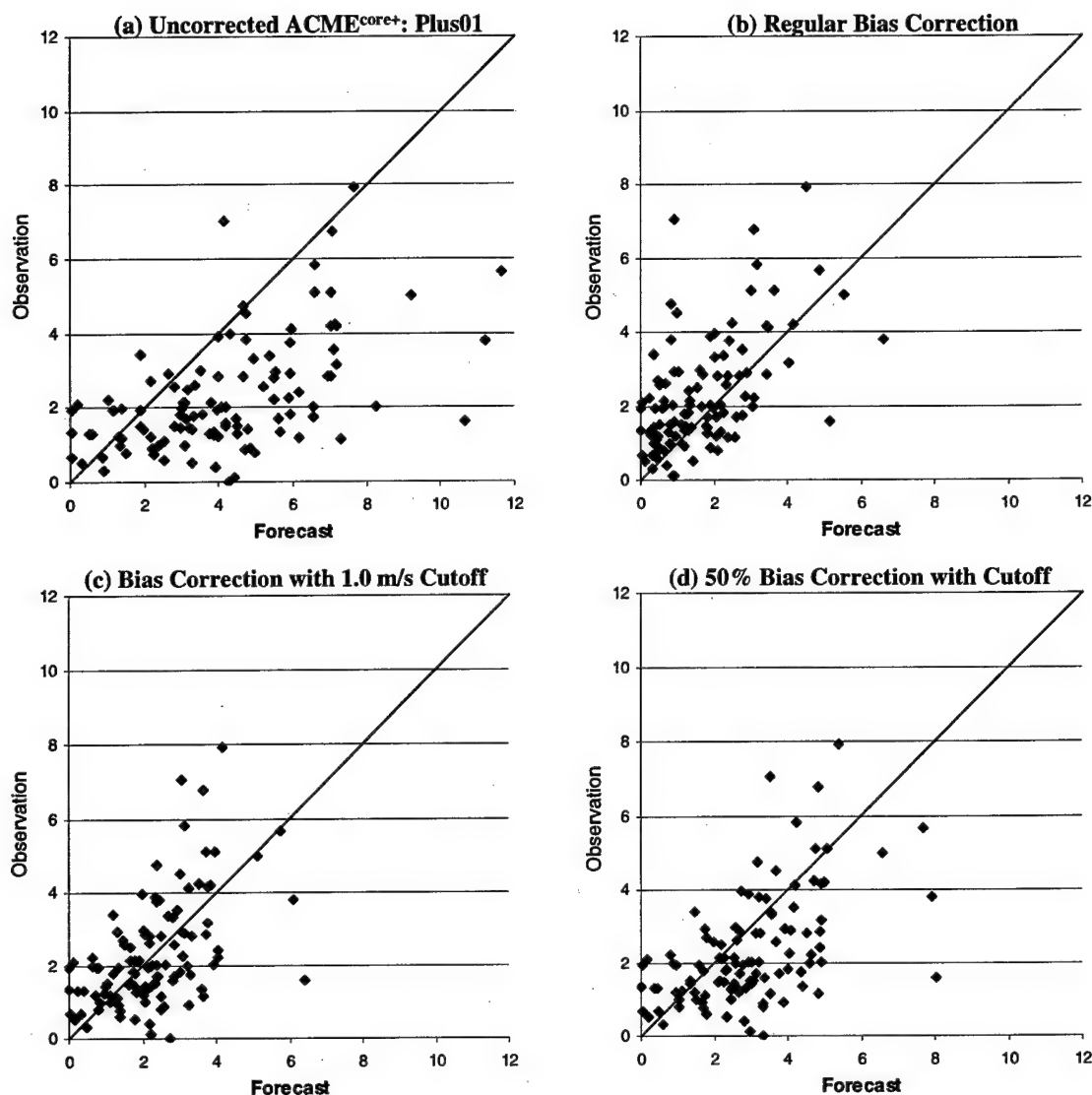


Figure 27. Scatter plots of 39-h forecast  $WS_{10}$  (in m/s) vs. RUC20 verification at a grid point in southern Puget Sound, WA, from member plus01 of ACME<sup>core+</sup>. (a) The uncorrected forecasts and observations showing the increase in error and variability of error with increasing wind speed, but with an obvious overforecasting bias. (b) Using the regular bias-correction method results in an overcorrection. (c) Using the regular bias-correction with a cutoff of 1.0 m/s (fcst. and obs. < 1.0 m/s are ignored), greatly improves the correction, but there are still too many underforecasts. (d) Reducing the multiplicative bias by 50% prevents the underforecasting problem.

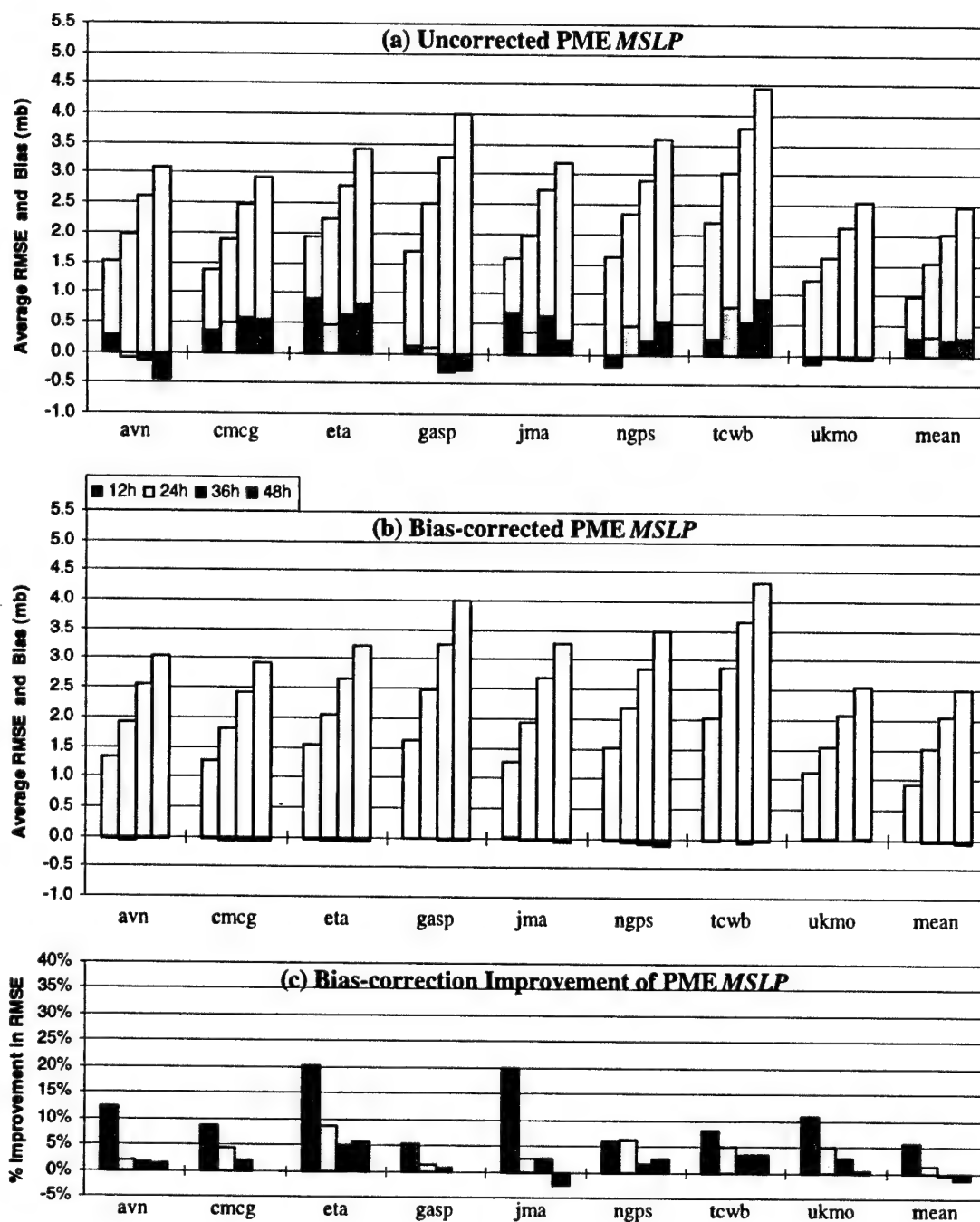


Figure 28. Results of *MSLP* bias correction for PME averaged over all bias-corrected cases, using the outer, 36-km domain. (a) and (b) show *RMSE* (clear histograms) and bias (shaded histograms) before and after bias correction, and the percent improvement (also shaded) in *RMSE* is given in (c). The results for the EF mean of PME is also shown.



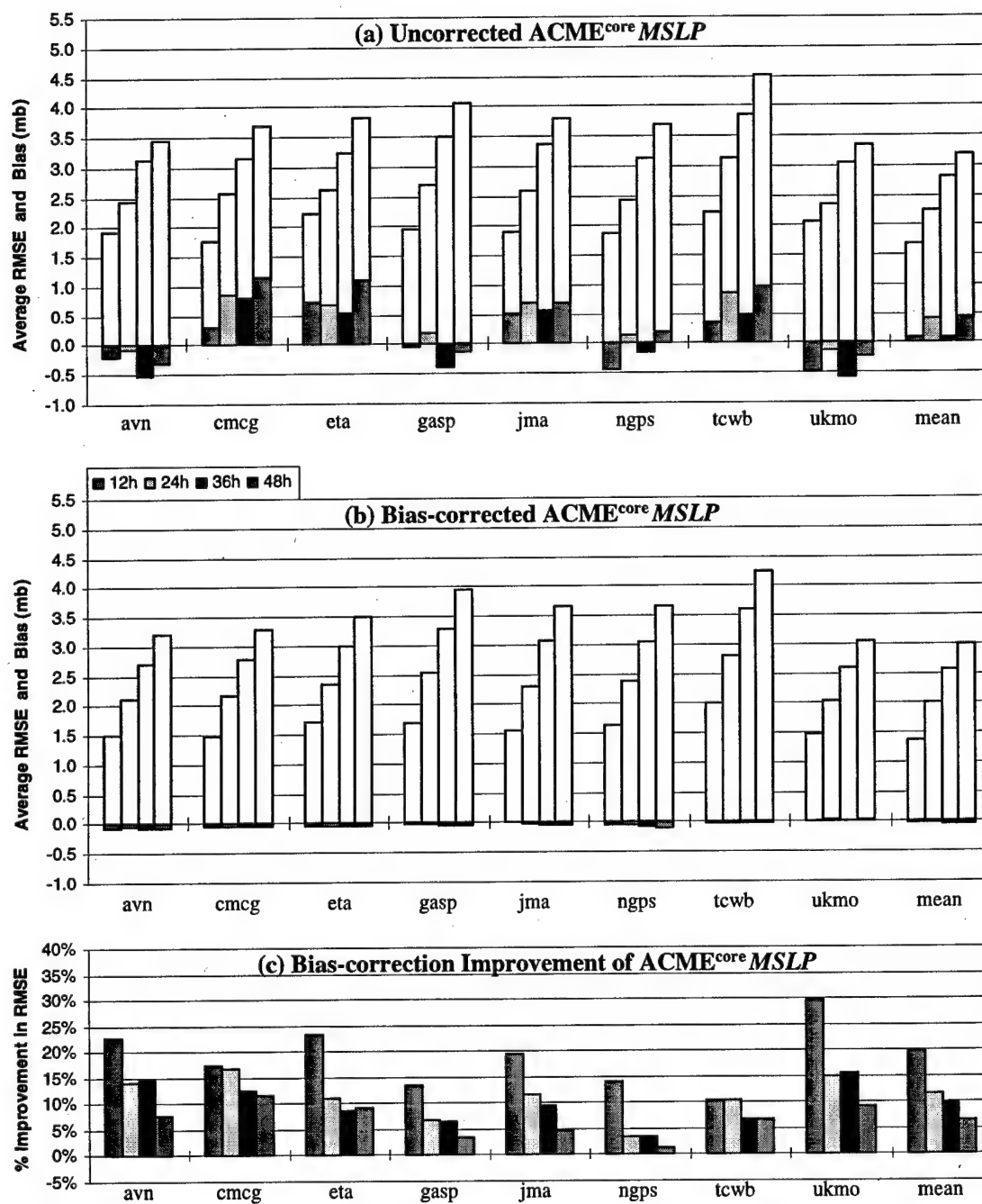


Figure 29. As in Figure 28 but for ACME<sup>core</sup>.

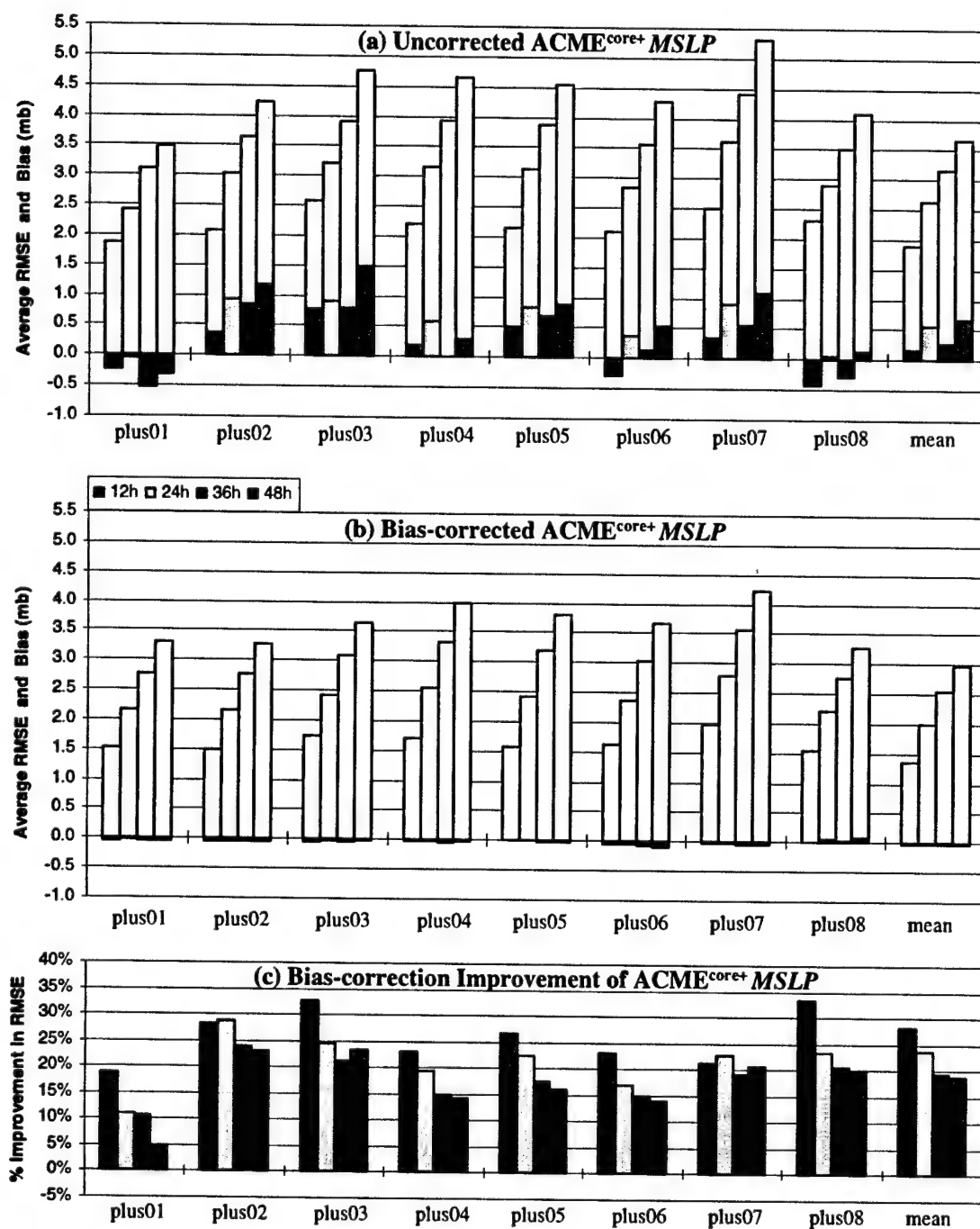


Figure 30. As in Figure 28 but for ACME<sup>core+</sup>.

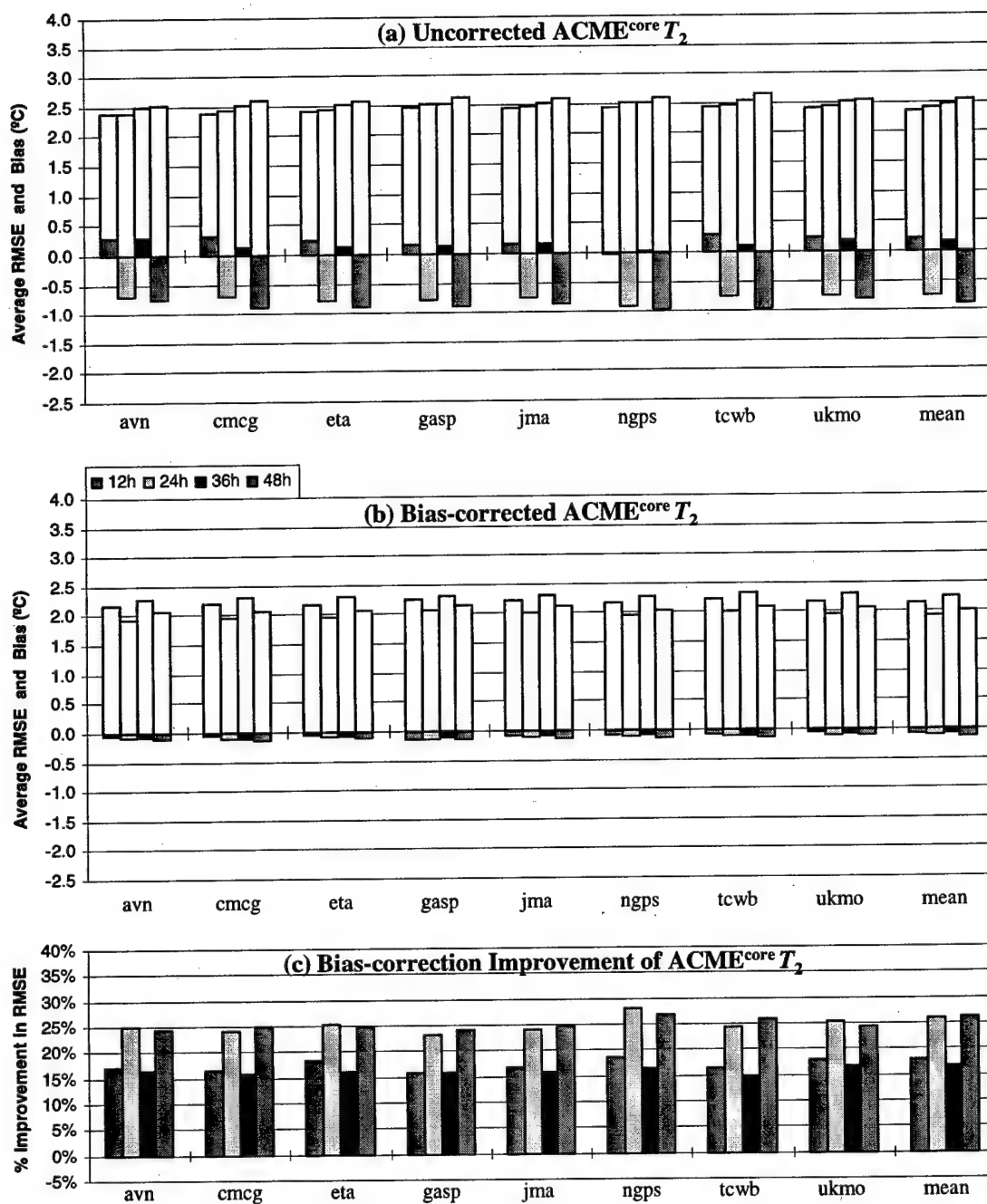


Figure 31. As in Figure 28 but for ACME<sup>core</sup>  $T_2$  data from the inner, 12-km domain, fit to the RUC20 20-km analysis grid.

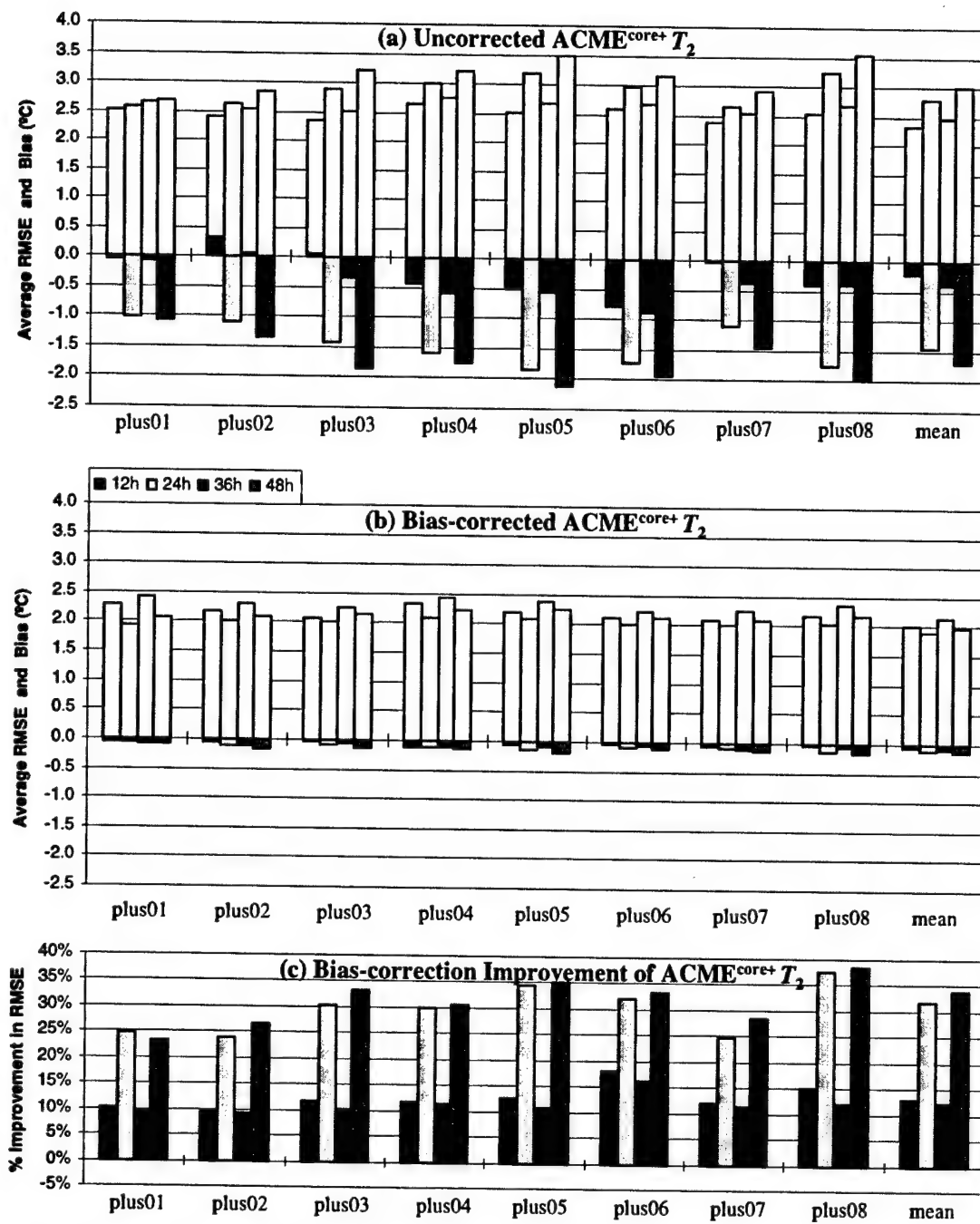


Figure 32. As in Figure 28 but for ACME<sup>core+</sup>  $T_2$  data from the inner, 12-km domain, fit to the RUC20 20-km analysis grid.

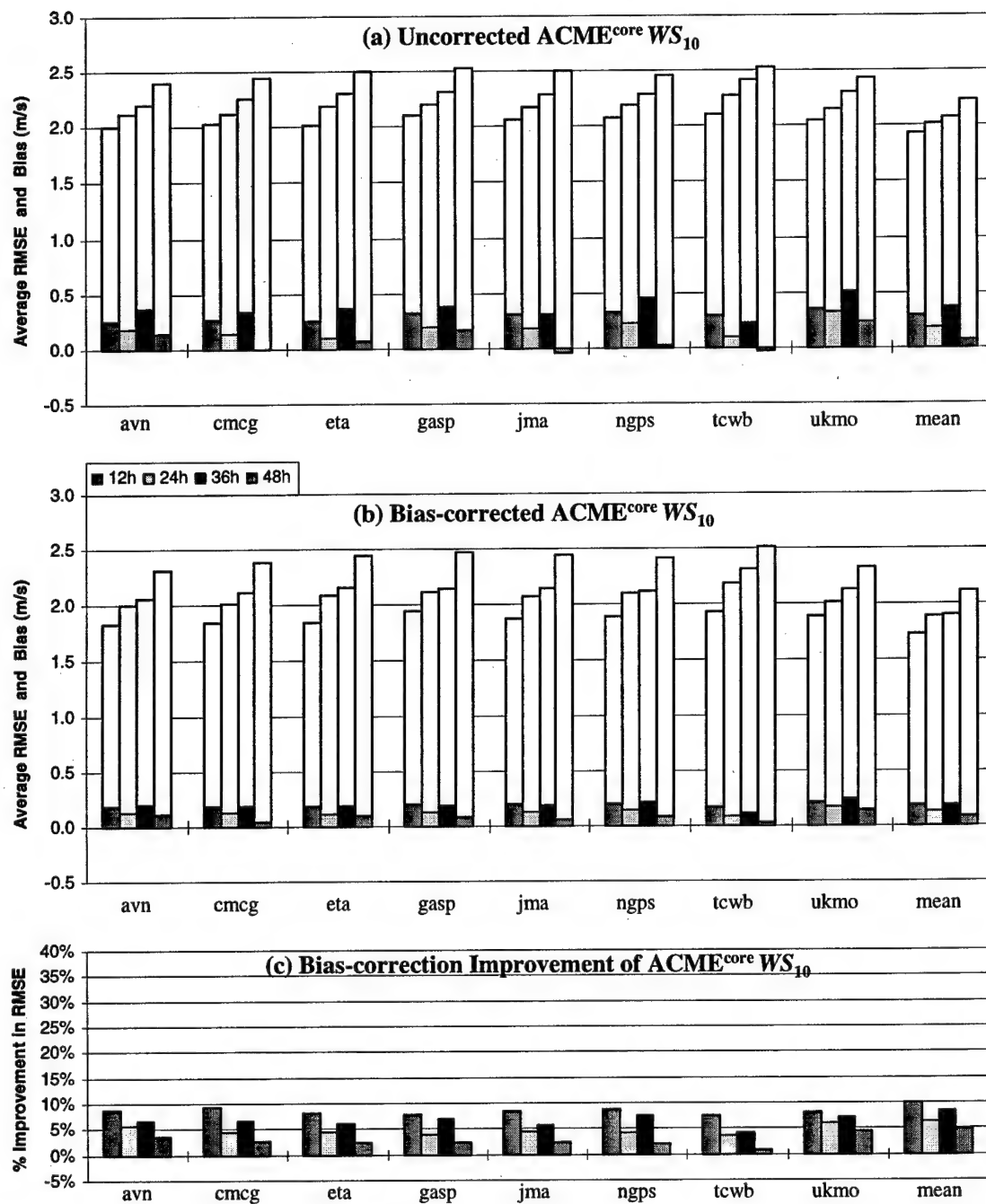


Figure 33. As in Figure 28 but for ACME<sup>core</sup> WS<sub>10</sub> data from the inner, 12-km domain, fit to the RUC20 20-km analysis grid.

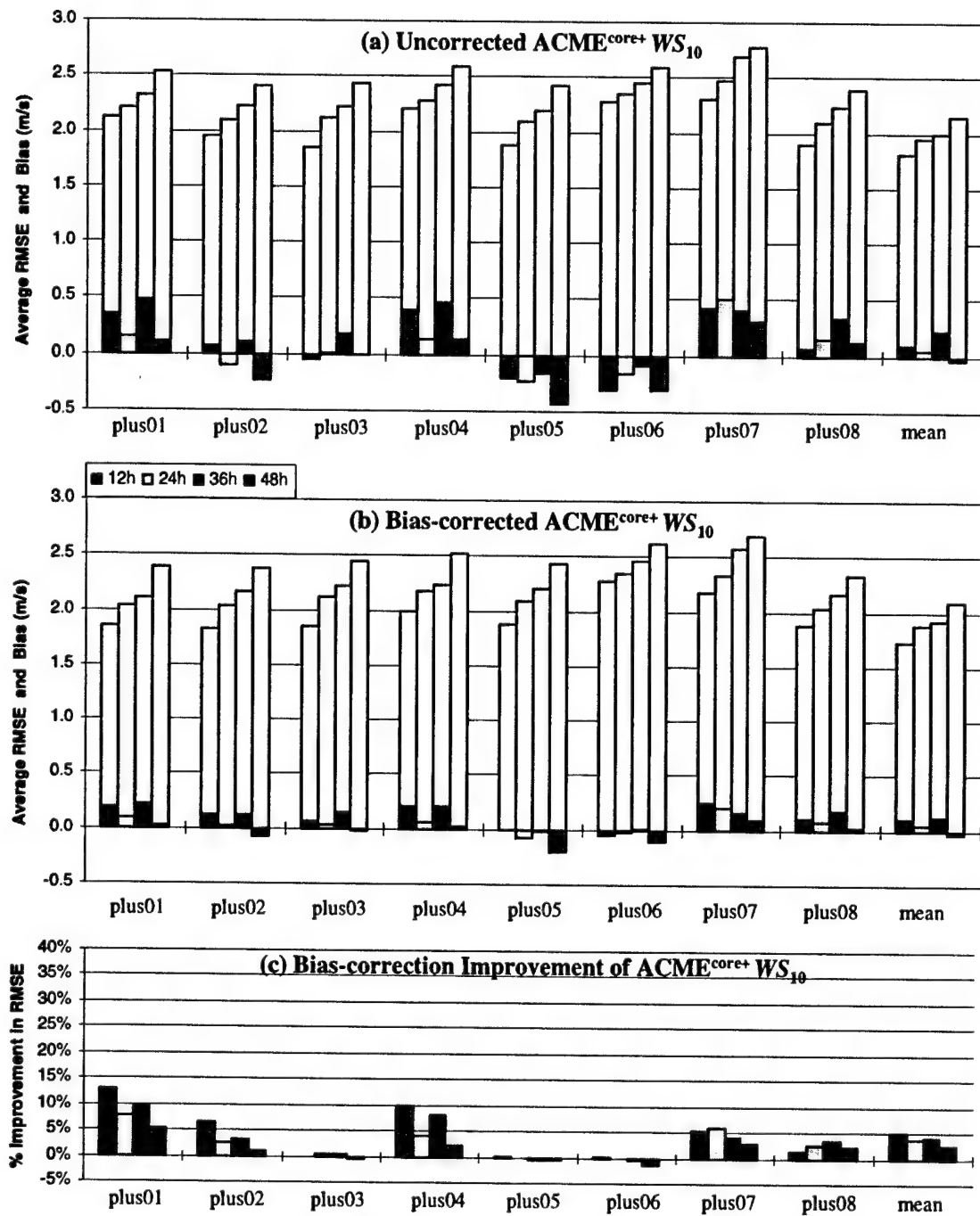


Figure 34. As in Figure 28 but for ACME<sup>core+</sup> WS<sub>10</sub> data from the inner, 12-km domain, fit to the RUC20 20-km analysis grid.

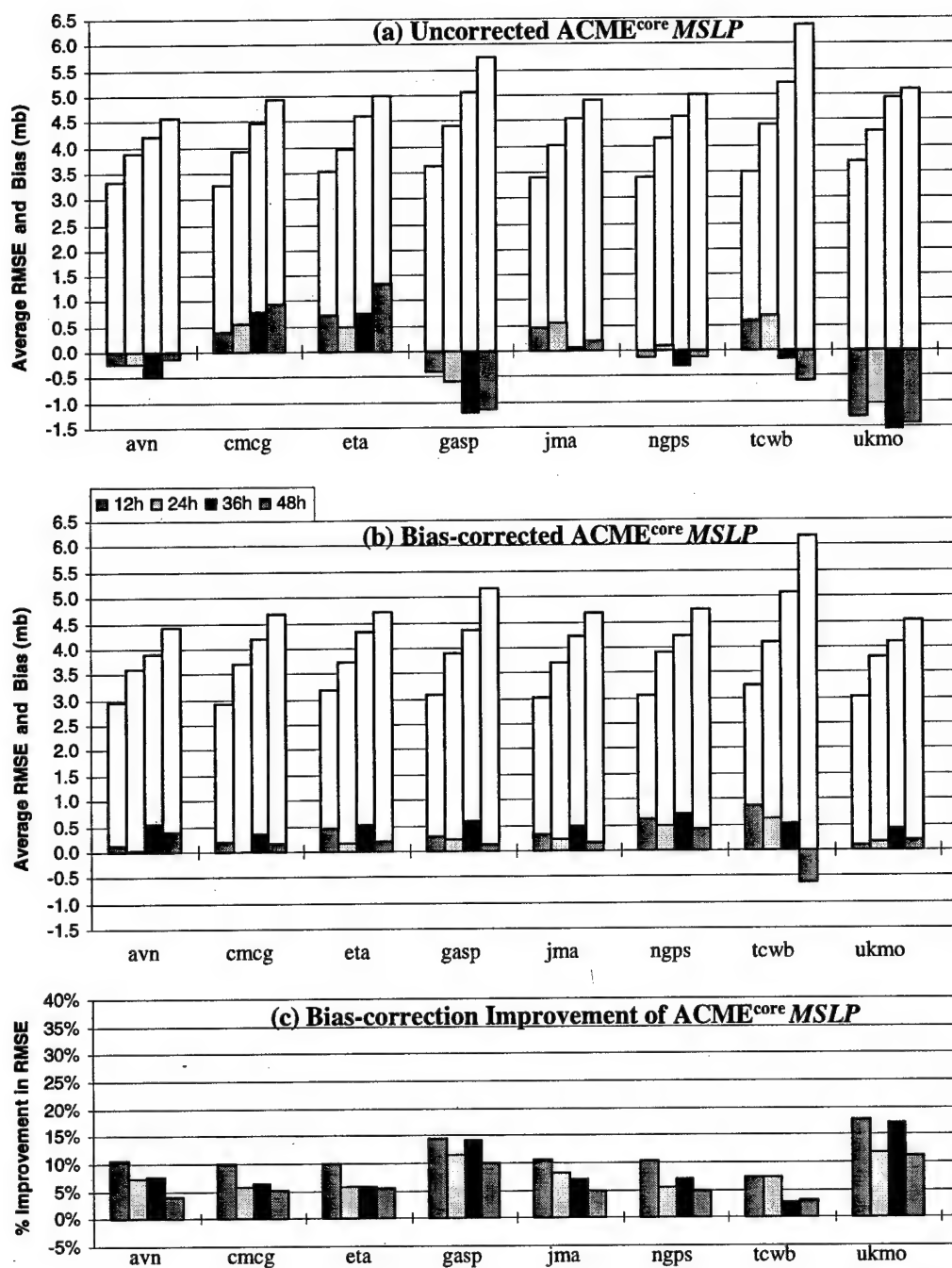


Figure 35. Observation-based verification results of *MSLP* bias correction for ACME<sup>core</sup> averaged over a one week period, using about 600 station observations in the outer, 36-km domain. The graphs are similar to the previous figures but note change in scale in (a) and (b) when comparing to Figure 29. Observation-based verification of the EF mean was not available.

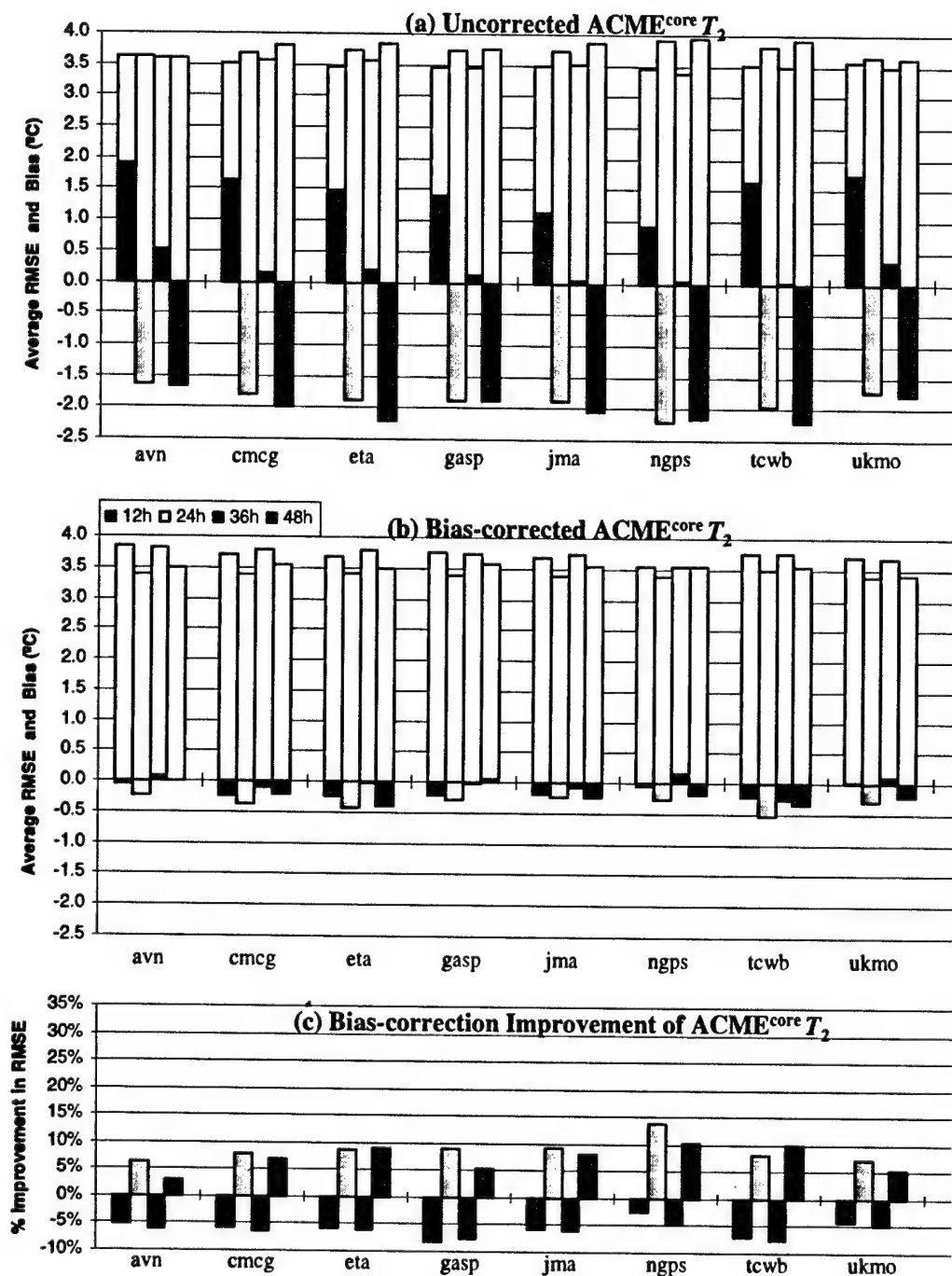


Figure 36. As in Figure 35 but for  $T_2$  data from the inner, 12-km domain. Note the downward-shifted scale in (c) when comparing to similar figures.



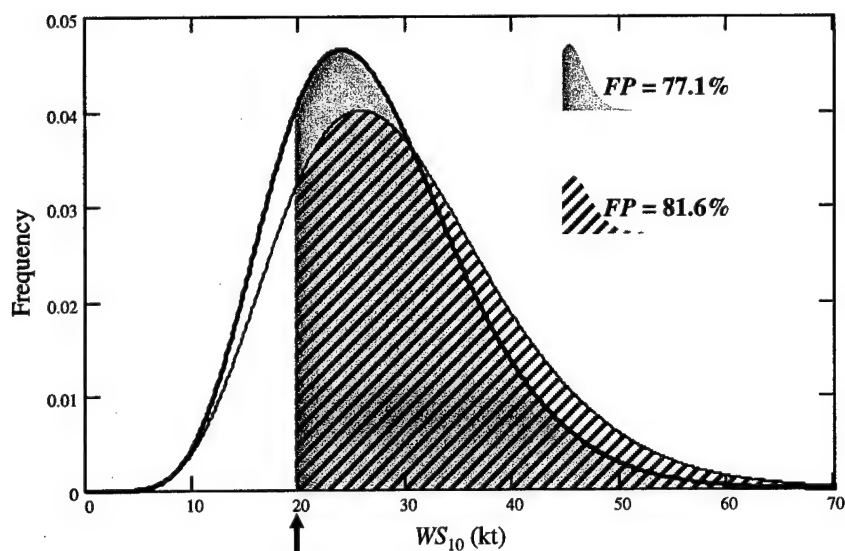


Figure 37. Example  $WS_{10}$  PDF at a model grid point. The thick curve is the PDF of an ideal ensemble with infinite members. The thin curve is a PDF fit to an ensemble of eight members (see text) drawn from the same ideal ensemble. The arrow indicates the event threshold ( $WS_{10} > 20$  kt) so the genuine  $FP$  for the event is the solid area under the thick PDF to the right of the event threshold. The hatched area is the 8-member ensemble's estimated  $FP$ .

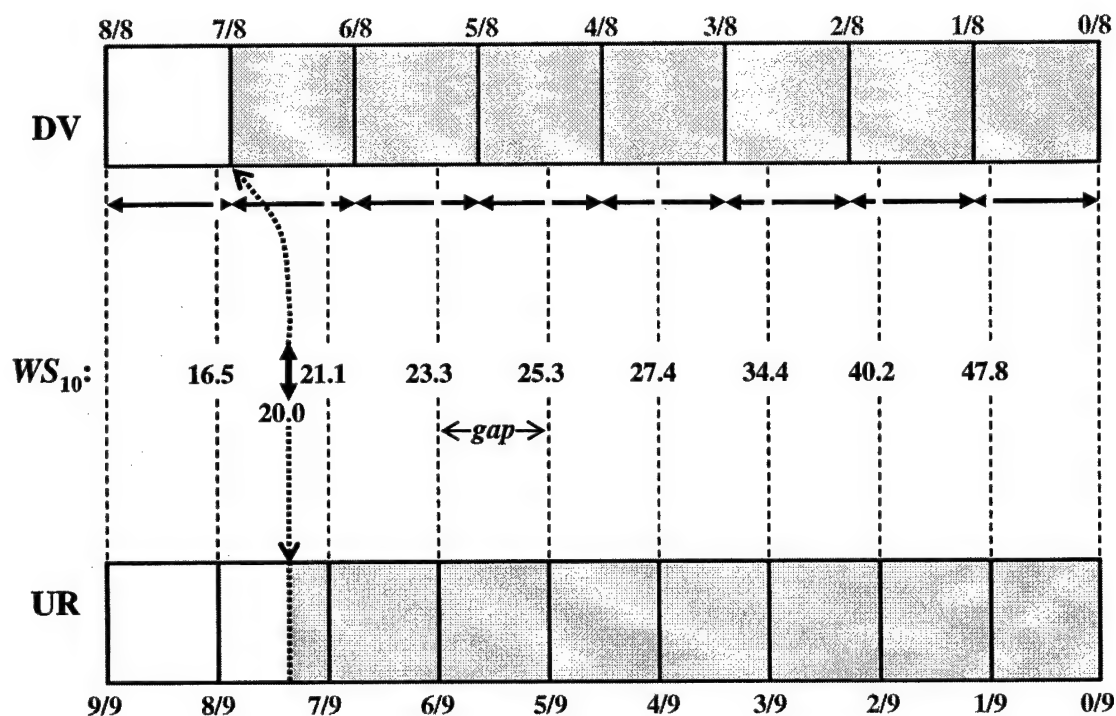


Figure 38. Schematic calculation of  $FP$  by DV and UR for the example ensemble  $WS_{10}$  forecast and an event threshold of 20.0 kt. A "gap" is the range of values between two ordered members.

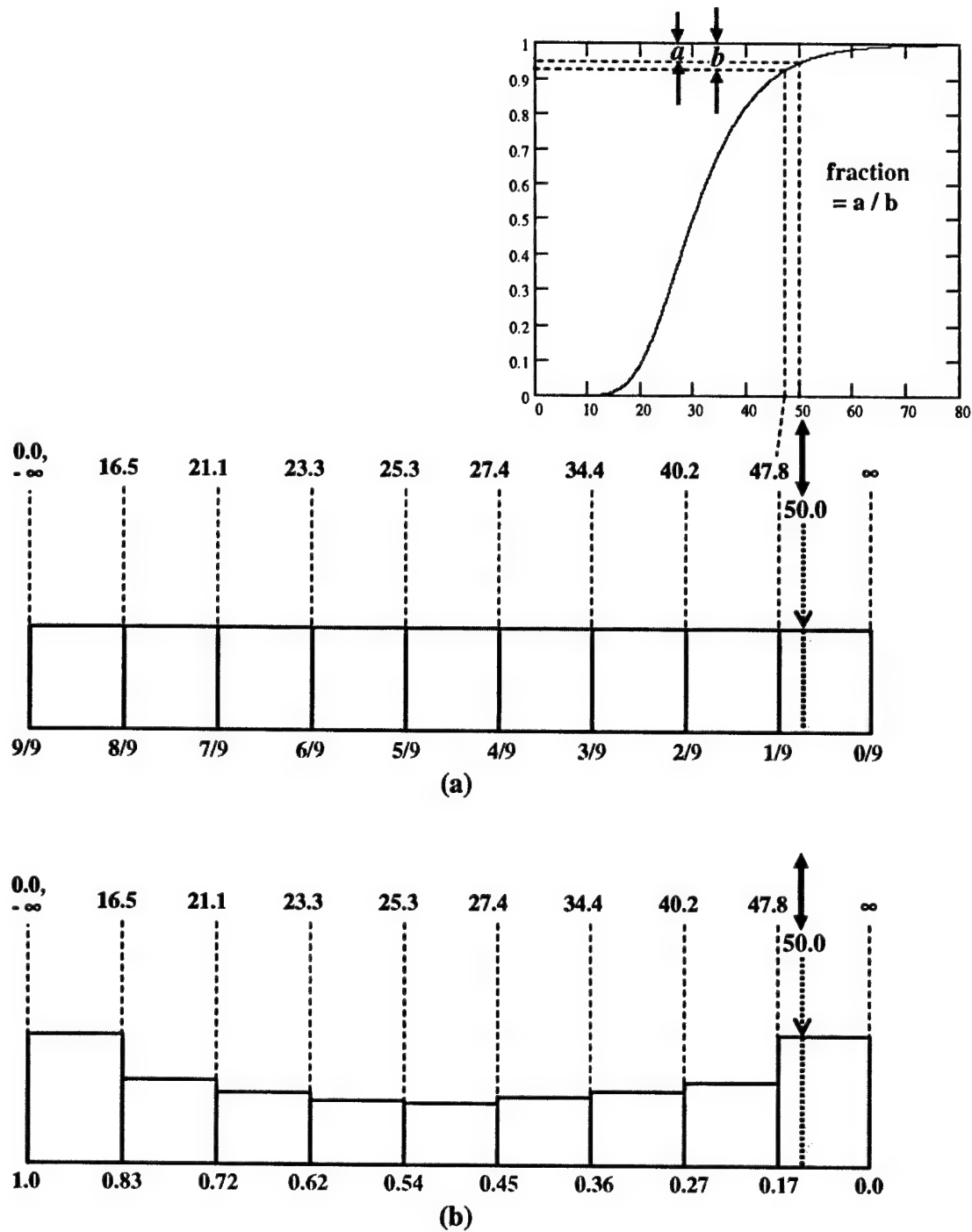


Figure 39. Schematic calculation of  $FP$  for an event threshold of 50.0 kt that occurs in the extreme right rank.  $FP$  is the shaded area, calculated by (a) UR, and by (b) weighted ranks for a hypothetical VRH.

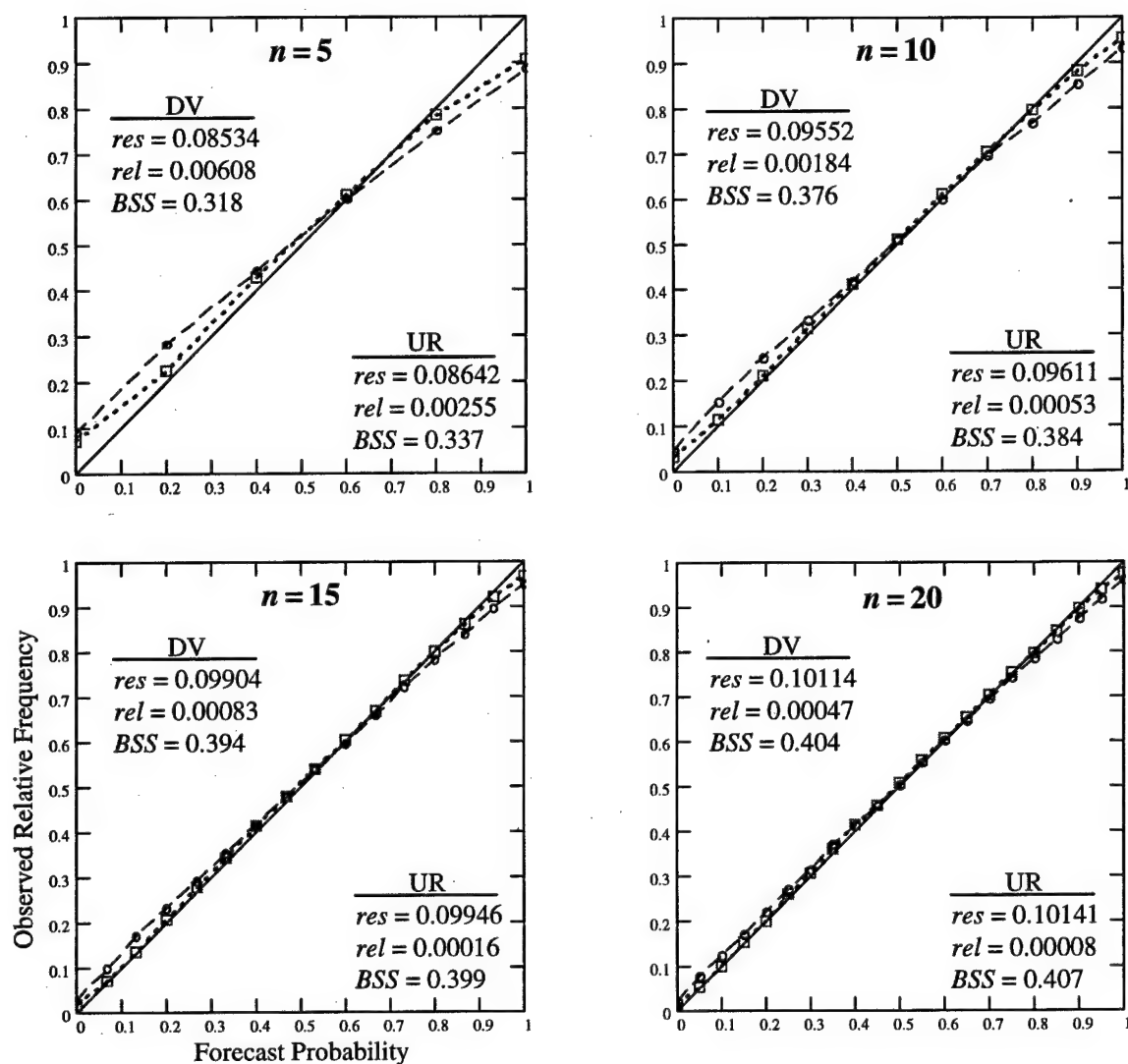


Figure 40. Effect of ensemble size and  $FP$  calculation methodology on  $FP$  skill. Results for  $FP$  calculated by DV are the dashed lines with circles. Results for  $FP$  calculated by UR are the dotted lines with squares. The solid diagonal is the line of perfect reliability which both calculation methods should produce since a perfect ensemble was simulated. To make the comparison fair, the continuous  $FP$  of UR was binned into the same number of bins (i.e.,  $n+1$ ) as set by the DV method. The Brier skill score ( $BSS$ ) and its components, resolution ( $res$ ) and reliability ( $rel$ ), are inset in each plot for the two methods.

Table 2. Brief description of the four SREF systems. (SMMA = Single-Model Multianalysis, PMMA = Perturbed-Model Multianalysis, and MMMA = Multimodel Multianalysis)

# of Mbrs.	Configuration Name	ICs	EF Type	Domain (km)	Forecast Interval (h)	Description
17	ACME	- 8 Analyses (core) - 1 Centroid - 8 Mirror	SMMA	36/12	3	- Analysis-Centroid Mirroring Ensemble - All members use the same version of MM5
	ACME <sup>core</sup>	8 Analyses (core)	SMMA	36/12	3	- Core subset of ACME - All members use same version of MM5
	ACME <sup>core+</sup>	8 Analyses (core)	PMMA	36/12	3	- Core subset of ACME - Each member has different version of MM5
8	PME	8 Analyses (core)	MMMA	36	6	- Poor Man's Ensemble - Each member has a different model (see Table 3)

Table 3. The eight analysis/forecast modeling systems of the PME, which also provide ICs for the ACME systems. (T, spectral truncation wave number; L, vertical levels; 3D-Var, 3-Dimensional Variational Data Assimilation; SSI, Spectral Statistical Interpolation; OI, Optimal Interpolation)









	Abbreviation/Model/Source	Type	Resolution (~ @ 45°N)		Objective Analysis
			Computational	Distributed	
	avn, Global Forecast System (GFS), National Centers for Environmental Prediction	Spectral	T254 / L64 ~55 km	1.0° / L14 ~80 km	3D-Var (SSI)
	cmcg, Global Environmental Multi-scale (GEM), Canadian Meteorological Centre	Finite Diff.	0.9°×0.9° / L28 ~70 km	1.25° / L11 ~100 km	3D-Var
	eta, Eta limited-area mesoscale model, National Centers for Environmental Prediction	Finite Diff.	32 km / L45	90 km / L37	3D-Var (SSI)
	gasp, Global Analysis and Prediction model, Australian Bureau of Meteorology	Spectral	T239 / L29 ~60 km	1.0° / L11 ~80 km	3D-Var
	jma, Global Spectral Model (GSM), Japan Meteorological Agency	Spectral	T106 / L21 ~135 km	1.25° / L13 ~100 km	OI
	ngps, Navy Operational Global Atmos. Pred. System, Fleet Numerical Meteorological & Oceanographic Ctr.	Spectral	T239 / L30 ~60 km	1.0° / L14 ~80 km	OI
	tcwb, Global Forecast System, Taiwan Central Weather Bureau	Spectral	T79 / L18 ~180 km	1.0° / L11 ~80 km	OI
	ukmo, Unified Model, United Kingdom Meteorological Office	Finite Diff.	5/6°×5/9° / L30 ~60km	same / L12	3D-Var



Table 5. Data of three core samples of  $n = 8$  from the normal distribution  $\mu = 5400$  gpm and  $\sigma = 15$  gpm, and the resulting centroid and mirror values.

Case	EF Group	Data	$\bar{x}$	$s$
(a)	Core	5398.3, 5399.7, 5394.5, 5407.3, 5417.0, 5411.9, 5418.7, 5409.1	5407.050	8.868
	A C M E	Centroid	5407.1	
		Mirrors	5415.8, 5414.4, 5419.6, 5406.8, 5397.1, 5402.2, 5395.4, 5405.0	8.295
(b)	Core	5412.6, 5413.7, 5385.4, 5417.9, 5399.1, 5397.8, 5372.0, 5400.3	5399.835	15.431
	A C M E	Centroid	5399.8	
		Mirrors	5387.1, 5386.0, 5414.3, 5381.8, 5400.6, 5401.9, 5427.6, 5399.4	14.434
(c)	Core	5367.0, 5383.2, 5408.9, 5395.8, 5404.5, 5424.3, 5396.1, 5364.6	5393.047	20.558
	A C M E	Centroid	5393.0	
		Mirrors	5419.1, 5402.9, 5377.2, 5390.3, 5381.6, 5361.8, 5390.0, 5421.5	19.231

Table 6. The standard MM5 land use table, in the exact file format employed in the MM5 code. The perturbed surface boundary parameters are albedo (ALBD, as a %), moisture availability (SLMO, as a fraction of 1.0), and roughness length (SFZO, cm). The other parameters include emissivity (SFEM, as a fraction of 1.0), thermal inertia (THERIN), snow-effect factor (SCFX), and heat capacity (SFHC).

USGS								
24,2,	'ALBD	SLMO	SFEM	SFZO	THERIN	SCFX	SFHC	
SUMMER								
1,	18.,	.10,	.88,	50.,	3.,	.52,	18.9e5,	'Urban and Built-Up Land'
2,	17.,	.30,	.92,	15.,	4.,	.60,	25.0e5,	'Dryland Cropland and Pasture'
3,	18.,	.50,	.92,	15.,	4.,	.60,	25.0e5,	'Irrigated Cropland and Pasture'
4,	18.,	.25,	.92,	15.,	4.,	.60,	25.0e5,	'Mixed Dryland/Irrigated Cropland and Pasture'
5,	18.,	.25,	.92,	14.,	4.,	.60,	25.0e5,	'Cropland/Grassland Mosaic'
6,	16.,	.35,	.93,	20.,	4.,	.60,	25.0e5,	'Cropland/Woodland Mosaic'
7,	19.,	.15,	.92,	12.,	3.,	.60,	20.8e5,	'Grassland'
8,	22.,	.10,	.88,	10.,	3.,	.62,	20.8e5,	'Shrubland'
9,	20.,	.15,	.90,	11.,	3.,	.60,	20.8e5,	'Mixed Shrubland/Grassland'
10,	20.,	.15,	.92,	15.,	3.,	0.,	25.0e5,	'Savanna'
11,	16.,	.30,	.93,	50.,	4.,	.56,	25.0e5,	'Deciduous Broadleaf Forest'
12,	14.,	.30,	.94,	50.,	4.,	.50,	25.0e5,	'Deciduous Needleleaf Forest'
13,	12.,	.50,	.95,	50.,	5.,	0.,	29.2e5,	'Evergreen Broadleaf Forest'
14,	12.,	.30,	.95,	50.,	4.,	.50,	29.2e5,	'Evergreen Needleleaf Forest'
15,	13.,	.30,	.94,	50.,	4.,	.54,	41.8e5,	'Mixed Forest'
16,	8.,	1.0,	.98,	0.01,	6.,	0.,	9.0e25,	'Water Bodies'
17,	14.,	.60,	.95,	20.,	6.,	.55,	29.2e5,	'Herbaceous Wetland'
18,	14.,	.35,	.95,	40.,	5.,	.58,	41.8e5,	'Wooded Wetland'
19,	25.,	.02,	.85,	10.,	2.,	.62,	12.0e5,	'Barren or Sparsely Vegetated'
20,	15.,	.50,	.92,	10.,	5.,	.60,	9.0e25,	'Herbaceous Tundra'
21,	15.,	.50,	.93,	30.,	5.,	.60,	9.0e25,	'Wooded Tundra'
22,	15.,	.50,	.92,	15.,	5.,	.60,	9.0e25,	'Mixed Tundra'
23,	25.,	.02,	.85,	10.,	2.,	.62,	12.0e5,	'Bare Ground Tundra'
24,	55.,	.95,	.95,	5.,	5.,	0.,	9.0e25,	'Snow or Ice'
WINTER								
1,	18.,	.10,	.88,	50.,	3.,	.52,	18.9e5,	'Urban and Built-Up Land'
2,	23.,	.60,	.92,	5.,	4.,	.60,	25.0e5,	'Dryland Cropland and Pasture'
3,	23.,	.50,	.92,	5.,	4.,	.60,	25.0e5,	'Irrigated Cropland and Pasture'
4,	23.,	.50,	.92,	5.,	4.,	.60,	25.0e5,	'Mixed Dryland/Irrigated Cropland and Pasture'
5,	23.,	.40,	.92,	5.,	4.,	.60,	25.0e5,	'Cropland/Grassland Mosaic'
6,	20.,	.60,	.93,	20.,	4.,	.60,	25.0e5,	'Cropland/Woodland Mosaic'
7,	23.,	.30,	.92,	10.,	4.,	.60,	20.8e5,	'Grassland'
8,	25.,	.20,	.88,	10.,	4.,	.62,	20.8e5,	'Shrubland'
9,	24.,	.25,	.90,	10.,	4.,	.60,	20.8e5,	'Mixed Shrubland/Grassland'
10,	20.,	.15,	.92,	15.,	3.,	0.,	25.0e5,	'Savanna'
11,	17.,	.60,	.93,	50.,	5.,	.56,	25.0e5,	'Deciduous Broadleaf Forest'
12,	15.,	.60,	.93,	50.,	5.,	.50,	25.0e5,	'Deciduous Needleleaf Forest'
13,	12.,	.50,	.95,	50.,	5.,	0.,	29.2e5,	'Evergreen Broadleaf Forest'
14,	12.,	.60,	.95,	50.,	5.,	.50,	29.2e5,	'Evergreen Needleleaf Forest'
15,	14.,	.60,	.94,	50.,	6.,	.58,	41.8e5,	'Mixed Forest'
16,	8.,	1.0,	.98,	0.01,	6.,	0.,	9.0e25,	'Water Bodies'
17,	14.,	.75,	.95,	20.,	6.,	.55,	29.2e5,	'Herbaceous Wetland'
18,	14.,	.70,	.95,	40.,	6.,	.58,	41.8e5,	'Wooded Wetland'
19,	25.,	.05,	.85,	10.,	2.,	.62,	12.0e5,	'Barren or Sparsely Vegetated'
20,	60.,	.90,	.92,	10.,	5.,	0.,	9.0e25,	'Herbaceous Tundra'
21,	50.,	.90,	.93,	30.,	5.,	0.,	9.0e25,	'Wooded Tundra'
22,	55.,	.90,	.92,	15.,	5.,	0.,	9.0e25,	'Mixed Tundra'
23,	70.,	.95,	.95,	5.,	5.,	0.,	12.0e5,	'Bare Ground Tundra'
24,	70.,	.95,	.95,	5.,	5.,	0.,	9.0e25,	'Snow or Ice'



### III. Results

In this chapter, we discuss the analysis of the four ensemble systems. Since each system has such unique attributes, intercomparison of the systems reveals answers to many questions surrounding SREF and indicates areas in need of further research. Appendix I reviews the various tools and metrics used in this chapter. Of particular note are two new measures—the standardized verification ( $V_Z$ ) and the verification outlier percentage ( $VOP$ ).

The results presented in this chapter may be influenced by the somewhat anomalous weather pattern of the 2002-2003 cool season. Figure 41 (borrowed from McMurdie and Mass, 2003) compares the  $Z_{500}$  mean and root-mean-square ( $RMS$ ) of the time-filtered  $Z_{500}$  for the past two cool seasons. Higher  $RMS$  values are indicative of larger  $Z_{500}$  variance and therefore more frequent storms. In a typical cool season, such as in Figure 41a, the Pacific NW experiences a fairly continuous train of extratropical cyclones from the predominantly zonal flow aloft over the eastern Pacific. In contrast, during the 2002-2003 cool season of Figure 41b, there were many prolonged periods of upper-level blocking that left the Pacific NW under a fair-weather ridge. Such blocking patterns are not unusual for the Pacific NW but are normally not so frequent. We suspect that the dominance of the fair-weather ridge pattern influenced our results, as we will describe later, but our general conclusions are not affected.

Unless otherwise noted, the analysis dataset was the 112 forecast cases of the bias-corrected subset (see Figure 10). Results using bias-corrected forecasts are denoted with an asterisk prior to the ensemble system's name (e.g., \*PME). The entire outer 36-km or inner 12-km grid domain was analyzed except for the outer most 5 rows and columns where lateral boundary condition (LBC) information was updated. All analysis of the predominantly synoptic-scale parameters, 500 mb geopotential height ( $Z_{500}$ ) and mean sea level pressure ( $MSLP$ ), was performed on the outer 36-km domain data using the centroid analysis (without tcwb) as verification. All analysis

of the mesoscale parameters, 2-m temperature ( $T_2$ ) and 10-m wind speed ( $WS_{10}$ ), was performed on the inner 12-km domain data using the RUC20 analysis as verification by fitting the MM5 12-km forecasts to the RUC20 20-km grid.

### A. Impact of Bias Correction

In the previous chapter, we demonstrated the need for and the positive results of our bias correction method from a primarily deterministic point of view and only touched on the possible impacts to SREF. Now we present results to describe two distinct benefits of removing bias in a SREF system: 1) that the quality of all SREF products, particularly forecast probability ( $FP$ ), is increased, and 2) that realistic evaluation and comparison of SREF systems is possible. The first benefit was anticipated but the improvement in SREF by bias correction exceeded expectations because the model biases are so large. The second benefit was unexpected, for as the analysis progressed, we discovered that only with bias-corrected data could we draw any firm conclusions.

From Figure 28 – Figure 34, it is clear that correcting bias in a SREF system reduces the  $MSE$  in each member and the ensemble mean. The benefit to EF is that  $FP$  skill is improved by approximately centering the forecast PDF on the mean of the verification's PDF so that the average error (verification – forecast) is close to zero. Recall from Figure 2 that a proper shift in the forecast PDF's location can adjust the  $FP$  toward the observed relative frequency ( $ORF$ ).

To explore improvement in SREF quality by bias-correction, Figure 42 displays a reliability diagram for ACME<sup>core+</sup>, before and after bias correction, in which the  $P(MSLP < 1001 \text{ mb})$  in the outer domain was forecast. This event threshold was chosen somewhat arbitrarily as the ~25<sup>th</sup> percentile of climatologic  $MSLP$  (Figure 1). While this event is not of direct concern in operational weather forecasting, it is worthwhile to analyze since  $MSLP$  is a common parameter among our SREF systems and important to forecasting in general. Think of the  $FP$  for this event as the chance of stormy weather (i.e., the probability of low  $MSLP$ ). In Figure 42, it is clear that

\*ACME<sup>core+</sup> provides far more reliable (i.e., closer to the diagonal) *FP* compared to ACME<sup>core+</sup>. To confirm this conclusion, Table 7 provides the data for calculating and plotting the reliability diagram. Notice that \*ACME<sup>core+</sup> also improved resolution and not just reliability, which will be discussed in detail below.

Figure 43 summarizes the reliability diagram results for *FP* of *MSLP* < 1001 mb at all lead times and for all our SREF systems except ACME. The plots of reliability (*rel*) and resolution (*res*) are on much different scales and the ordinate axis of the reliability plot is reversed so that upward is better on all plots. Recall that  $BSS = (res - rel) / unc$ , where *BSS* is the Brier skill score and *unc* is the uncertainty term.

The first thing to notice in Figure 43 is that all the SREF systems are highly skilled in the short range at forecasting this event (i.e., *BSS* far above 0.0), which should be expected for a predominantly synoptic-scale parameter such as *MSLP*. The improvement in *BSS* by the bias correction is relatively large (~3%) for the ACME systems but insignificant for PME since the PME members displayed much less bias compared to ACME systems, as discussed in Chapter II. To more intuitively quantify the significance of the ~3% improvement in *BSS* for the ACME systems, we can examine the skills of the systems before and after bias correction across lead time. On average, the skill of \*ACME<sup>core</sup> or \*ACME<sup>core+</sup> is the same as that of the uncorrected systems six hour previously. In other words, there was a six-hour improvement in *FP* skill by the bias correction—a great improvement in the short-range. Lastly, note that there are roughly equal contributions from both reliability and resolution to the *BSS* improvement by bias correction, which holds true for all parameters and events that we examined.

Figure 44 provides the *BSS* results for forecasts of  $P(T_2 < 0^\circ\text{C})$  in the inner domain—a more operationally significant event and one intimately connected with model physics and surface boundary parameterizations (SBPs). The *BSS* improvement by bias correction is about twice as

large as that for *MSLP* because, as we saw in the last chapter,  $T_2$  has a much more pronounced bias. There is also a diurnal cycle in *BSS* that appears contrary to *RMSE* of bias-corrected  $T_2$  (Figure 31) in which the late afternoon (i.e., 24- and 48-h lead times)  $T_2$  has lower *RMSE* and should therefore correspond with higher *FP* skill rather than lower as in Figure 44. The sharp dip in uncorrected *FP* skill in the late afternoon is due, in part, to the extreme bias during that period. After bias correction, the late afternoon reliability is on a par with the other times of the day but the dip remains in the *BSS* due to the resolution fluctuation. The marked diurnal signal in the resolution is tied to the variability in uncertainty. Intuitively and mathematically, one would expect more skillful *FP* in the late afternoon when uncertainty is at a minimum with  $T_2 < 0^\circ\text{C}$  occurring less often (i.e., lower sample climatology, *SC*). However, an event that occurs less often in space or time is more difficult to discriminate, thus the drop in late afternoon resolution and lower *BSS*.

The increase in resolution by the bias correction is an important finding that indicates a sharpening of the forecast PDF, or a reduced variance among the ensemble members. Referring back to Figure 37, one can imagine that for any given event threshold, a more narrow PDF is more likely to produce *FP* toward the extreme values (i.e., 0% and 100%), which of course increases *res*. Table 7 shows that the bulk of the better resolution of  $^*\text{ACME}^{\text{core+}}$  compared to  $\text{ACME}^{\text{core+}}$  came from a 17% increase in the number of forecasts in the 100% *FP* bin. There is actually a ~1% decreased weighting in the lower extreme *FP* by  $^*\text{ACME}^{\text{core+}}$  because the bias correction shifted the PDF to the right as well as reduced the spread.

The reason for the reduced spread is that the bias correction adjusts each ensemble member toward the verification (gridded analysis) differently. The members have different biases, both in magnitude and direction from the verification, but are all corrected toward a common center, thus reducing variance. Figure 45 shows evidence of the decrease in EF spread by bias correction,

which should be expected if EF spread is a reflection of uncertainty. In effect, removing model bias in an ensemble eliminates bogus uncertainty in the sense that there is no uncertainty in systematic errors that can be identified and corrected. Before systematic errors are removed, they appear to be part of the uncertainty since they contribute to the forecast error. After systematic errors are removed, the stochastic error remains as the true uncertainty, which can not be corrected but may be accounted for with a well-formulated ensemble system.

Besides significantly improved *FP*, our second point about the importance of bias removal is that it allows for realistic evaluation and comparison of SREF systems. Consider trying to determine if  $ACME^{core+}$  provides benefit over  $ACME^{core}$  in Figure 44.  $ACME^{core+}$  performs better than  $ACME^{core}$  at some lead times (e.g., 6 h – 15 h) and the same or worse at other lead times (e.g., 18 h – 24 h). It is only after bias correction that  $*ACME^{core+}$  clearly stands out as superior to  $*ACME^{core}$ .

As another example of bias-removal benefit to analysis, Figure 46 shows verification rank histograms (VRHs) of *MSLP* for PME,  $ACME^{core}$ , and  $ACME^{core+}$  before and after bias correction. Forecasts with a significant and consistent bias cause a shift of the rank probability toward one side. Notice that in the PME VRH the overdispersion is much more evident after the bias removal. A more disconcerting problem comes about if forecasts have a dual bias that changes over time, resulting in a strongly u-shaped histogram which may lead to an incorrect conclusion that the ensemble is underdispersive (Hamill, 2001). Removing the bias by a method such as ours eliminates that possibility.

As a final comment on Figure 46, notice that bias correction barely altered the verification outlier percentage (*VOP*) for PME and  $ACME^{core}$ , but *VOP* was improved for  $ACME^{core+}$ . One could attribute a lower *VOP* to an increase in ensemble spread, which may allow truth to be portrayed more often. However, as discussed above, bias correction decreases ensemble spread,

especially for ACME<sup>core+</sup> in which there is more variability among the model biases (i.e., compare Figure 31 with Figure 32). We conclude then that bias correction does not cause better portrayal of truth by simply adding unrealistic spread but by shifting the PDF toward regions of verification values previously not portrayed. The additional benefit of increased resolution from a sharper forecast PDF occurs simultaneously.

## B. Model Uncertainty

### 1. Multimodel vs. Perturbed-model

This section addresses the relative merits of the multimodel and perturbed-model approaches for accounting for model uncertainty by comparing the results of the PME to the ACME<sup>core+</sup> system. This comparison is restricted to the 36-km domain because the PME consists of only large-scale models with coarse grids. Additionally, the PME does not contain many of the surface forecast parameters of interest (e.g.,  $WS_{10}$  and  $T_2$ ) so only  $Z_{500}$  and  $MSLP$  were considered in this analysis.

#### a) Dispersion

We begin by examining dispersion diagrams (Figure 47a & b) to explore the systems' ability to represent forecast uncertainty. Recall that the dispersion diagram is like an error variance diagram except that the plotted curves are the EF spread (i.e., variance of ensemble members) and the  $MSE$  of the EF mean, which are required to match for statistical consistency after adjusting for ensemble size as in Equations (7) and (8). The  $MSE$  of the EF mean should be thought of as the 'target variance' that an ensemble should have to properly represent forecast uncertainty. Since each of our SREF systems has a different EF mean  $MSE$ , each system should be plotted on a separate diagram to avoid confusion. For example, the plots in Figure 45 contain results for

both PME and ACME<sup>core+</sup> in which the higher spread of PME suggests improved statistical consistency by PME since ACME<sup>core+</sup> is underdispersive (shown later). However, it is meaningless to directly compare the EF spread of the two systems since the target variance is so different between the systems. An exception is that ACME<sup>core</sup> and ACME<sup>core+</sup> have such a similar target variance (i.e., similar EF mean *MSE*) that it is instructive to plot and analyze them together (covered in the next section).

Before interpreting Figure 47a & b, there are a few more things to note:

- 1) The apparent decrease in EF spread from 0 to 12 h is due to both the bias correction (which was not applied at the 0 h) and the MM5 spin-up period. Figure 45a shows that without bias correction the error growth is large in the first 12 h for PME, but only slight for the ACME systems. In the early part of forecast integration, the MM5 adjusts the information from the large-scale models to fit the MM5 attractor so the ACME systems' solutions become more similar and error growth is restricted.
- 2) The 12-h *MSE* is likely an underestimate of the actual error since the verification (centroid analysis) contains much of the forecast information due to use of the forecasts as first guess fields in the objective analysis routines of the core analyses. By the 24-h lead time and beyond, the centroid analysis can be considered independent of the forecasts.
- 3) For reference, we included the climatic variance ( $\sigma_c^2$ ) to show how far below error saturation the results are in the short range. The  $\sigma_c^2$  values were found using all the verification data for the full dataset from the avn analysis for  $Z_{500}$  and *MSLP* (e.g., see Figure 1a), and from the RUC20 analysis for  $WS_{10}$  and  $T_2$ . As a confirmation, our  $Z_{500}$   $\sigma_c^2$  of 14,500  $\text{gpm}^2$  is comparable to what is shown in Figure 3b.

The most striking difference between PME and ACME<sup>core+</sup> in Figure 47 is that the PME is slightly overdispersive while the ACME<sup>core+</sup> is very underdispersive. The forecast PDFs

produced by the PME are evidently too wide compared to the PDF from which truth is drawn so the PME identifies more uncertainty than is actually present—highly unusual for an EF system. ACME<sup>core+</sup> shows the more typical result of an EF system producing too narrow a forecast PDF, and failing to represent all the uncertainty.

Since PME and ACME<sup>core+</sup> used the same ICs, the difference in their dispersive characteristics likely reveals that the multimodel system (PME) is able to more accurately represent model uncertainty compared to the perturbed-model system (ACME<sup>core+</sup>). We expect the PME to exhibit greater dispersion since it has more model diversity, but the overdispersion of PME may mean that the model differences among the PME members are too great. Regardless, the severity of PME's overdispersion is much less than the large underdispersion of ACME<sup>core+</sup>. It appears that even with the extensive efforts in building model diversity into ACME<sup>core+</sup>, the perturbed-model approach does not represent many critical aspects of model uncertainty that the multimodel approach can, such as the model numerics.

Examining  $Z_{500}$  and *MSLP* VRHs in Figure 48a & b confirms the results of the dispersion diagrams and provides more details. For these synoptic parameters, \*ACME<sup>core+</sup> performed well (nearly uniform VRHs), but it is clear that \*PME was more successful at portraying truth. The *VOP* scores show that for *MSLP*, truth was not portrayed 1.55% of the time by \*PME vs. 6.67% of the time by \*ACME<sup>core+</sup>. The slight overdispersion of \*PME is evident in the subtle n-shape of the \*PME VRHs.

While the superior statistical consistency of \*PME over \*ACME<sup>core+</sup> is unquestionable, there are other possible reasons for the difference between the two systems besides the systems' relative ability to represent model uncertainty:

- 1) The coarse grid resolution of the PME members (c.f. Table 3) may account for part of the PME's lower *MSE*, which makes the PME appear more statistically consistent. However,



since only the 36-km domain was considered in this analysis, the effect of resolution is likely minor. Also, some PME members have fairly high resolution (i.e., eta at 32 km and avn at ~55 km).

- 2) As discussed above, there is a spin-up period in the  $ACME^{core+}$  solutions that restricts error growth early in the forecast cycle. While the spin-up effect does account for some of the low spread of  $*ACME^{core+}$ , Figure 49a shows that the lower dispersion of  $*ACME^{core+}$  is not due to the spin-up effect. Once the spread of both systems is matched at the 12-h lead time (well after spin-up),  $*PME$  clearly shows more dispersion than  $*ACME^{core+}$ .
- 3) It has been shown that an ensemble that uses a limited-area model (LAM) has lower dispersion compared to an ensemble that uses a much larger model domain (Nutter, 2003). Beyond the issue raised by Errico and Baumhefner (1987) who pointed out that when using a LAM, the LBCs as well as the ICs must be perturbed to avoid limiting predictability error growth, Nutter (2003) described how the use of periodically updated LBCs may act to filter out short waves and reduce nonstationary wave amplitude from the large-domain model providing the LBCs to the LAM. This effect can cause errors in the LAM solution, but more importantly, it may cause an ensemble of LAM solutions to share similar errors even when they have different perturbed LBCs (as in  $ACME^{core+}$ ), thus reducing spread and causing underdispersion during the forecast cycle.

To explore how much of the weaker dispersion of  $ACME^{core+}$  may be from filtering of waves in the LBCs versus use of incomplete representation of model diversity, we can examine plots of standardized verification ( $V_Z$ ). In Figure 50  $Z_{500} V_Z$  is plotted using bias-corrected forecasts since the biased forecasts have unrealistically high standard deviation and thus  $V_Z$  values that are too small.  $Z_{500}$  was used so that synoptic-scale wave effects could be studied. The forecast case in

Figure 50 is one with above average *VOP* in which truth really got away from the ensemble. Forecast cases with an average or low *VOP* level had similar results but were not as definitive.

As shown in Figure 50, at the 12-h lead time there was long wave trough along the West Coast with a rapidly approaching short wave around 150°W that initialized near the boundary (not shown). As the short wave dove into the long wave trough at 24 h and 36 h, the high  $V_z$  values reveal that in this region truth evolved quite differently compared to all the \*PME members and much more so for \*ACME<sup>core+</sup> members. The EF mean is included in these plots to show the solution about which the members are varying. Looking at both the EF mean and  $V_z$ , it appears that the \*ACME<sup>core+</sup> members are too clustered about a solution with a slower and deeper short wave off the Pacific NW coast and thus failing to portray the truth. By the 48-h lead time, the \*PME is portraying truth fairly well but the \*ACME<sup>core+</sup> members continued to stay clustered together with solutions much different from the truth.

How much of the truth not portrayed by \*ACME<sup>core+</sup> is due to weak model diversity and how much from use of a LAM? The short wave analyzed above was initialized partly within the domain so it may have suffered some filtering as it completed its entry through the lateral boundary. Additionally, effects of downstream development could have played a role so that filtered waves entering later produced further limits to the \*ACME<sup>core+</sup> error growth about the wave of interest. The large-scale models of the PME were able to more accurately develop the wave as well as represent more likely possibilities (i.e., higher and meaningful dispersion) since waves on all scales are represented over a much larger domain. It appears possible that some of the higher *VOP* of \*ACME<sup>core+</sup> is due to filtering information in the LBCs, which makes \*ACME<sup>core+</sup> members share similar errors and reduces dispersion.

However, there is also strong evidence in Figure 50 that it is the weak model diversity of \*ACME<sup>core+</sup> (relative to \*PME) that increases the *VOP*. Consider again the 12-h lead time in

Figure 50 in which \*ACME<sup>core+</sup> developed an area of truth not portrayed on the Oregon coast and greatly expanded the area over British Columbia compared to \*PME. At only 12 h into the forecast cycle and thousands of kilometers from the lateral boundaries, it is highly unlikely that these problems were caused by LBC wave filtering but rather were due to the weak model diversity of ACME<sup>core+</sup>. Furthermore, such evidence of the large impact by the weak model diversity leads us to speculate that model diversity was also a large factor (perhaps larger than the LAM effect) in the above analysis of the short wave that caused so much trouble in Figure 50.

We can not make firm conclusions regarding the relative contribution of LAM and model diversity effects on the different dispersions of PME and ACME<sup>core+</sup> since it is extremely difficult to separate out the two effects. However, the analysis results and consideration of the design of the two systems (multimodel vs. perturbed-model) suggest that the lower dispersion of ACME<sup>core+</sup> is primarily a result of its inability to capture the amount of meaningful model uncertainty that is captured in PME. As an aside, notice in Figure 50 that much of the difficulty with portraying truth originates from the core analyses and their forecasts (i.e., PME) and are amplified in ACME<sup>core+</sup>. We will discuss this further below when covering the performance of ACME.

#### b) Skill and Utility

Returning to Figure 43, consider the *FP* skill of the multimodel and perturbed-model approaches. Measuring skill improvement again by forecast lead time, \*PME outperformed \*ACME<sup>core+</sup> by about 11 h. Even though the multimodel approach overrepresents uncertainty, it yields far superior results to the perturbed-model approach that grossly underrepresents uncertainty. Note that the higher *BSS* superiority of \*PME is completely due to better resolution. The reliability of \*PME is basically the same or slightly lower compared to \*ACME<sup>core+</sup>, but this does not mean increased model diversity cannot improve reliability. The lack of difference in

reliability between the two systems is simply because both systems are nearly perfectly reliable for this event in the short range.

There is an apparent contradiction concerning the resolution improvement and the increased model diversity of \*PME. In analyzing the impact of bias removal, we discussed that the decreased spread (narrowing of the forecast PDF) improves resolution, and in the last section we demonstrated that model diversity increases spread (widening the forecast PDF). So how can \*PME improve resolution if it has greater spread?

The resolution improvement by \*PME can be diagnosed by comparing reliability diagram results of \*ACME<sup>core+</sup> and \*PME (Figure 42 and Table 7). Resolution can be improved (i.e., higher *res* value) in two different ways:

- 1) Increase the weight (i.e., number of forecasts) in the *FP* bins toward the *FP* extremes.
- 2) Regardless of the weight in each bin, shift points on the reliability diagram toward the *FP* extremes (i.e., further from the zero skill line)

By either means, an ensemble is then better at discriminating between whether an event will occur or not. \*PME has better resolution than \*ACME<sup>core+</sup> only because the \*PME points are shifted toward the *FP* extremes. In fact, \*PME actually negated some its resolution improvement with reduced weighting of the *FP* extremes (i.e., in Table 7 compare # of forecasts for *FP* of 0% and 100% between \*ACME<sup>core+</sup> and \*PME).

These results allow us to resolve the apparent contradiction of \*PME's improved resolution and higher spread. Lower (higher) EF spread always improves (worsens) resolution, but resolution can also be improved in a more subtle way, which can be imagined using Figure 37 once again. Say the event threshold is at the right end of the \*ACME<sup>core+</sup> PDF so that *FP* = 0%. The \*PME PDF extends out a bit further, representing more possible values where the verification will likely occur, and so \*PME may give a slightly higher *FP*, say 3%. For such low

*FP*, the verification will rarely occur above the event threshold, but if the increased spread of \*PME does give real possibilities, then the event will eventually occur and confirm the validity of the \*PME *FP*. A wider forecast PDF can therefore provide improved *res* if it better represents the possible occurrence of the event. Consider an EF system in which one member's PBL scheme is switched to something totally different from the other members, but equally valid. This perturbed member may give a  $T_2$  value completely outside of the other members, increasing EF spread and identifying a possible value of truth not previously part of the ensemble. The new ensemble of  $T_2$  values is better at discriminating the event since it identifies more valid possibilities. This effect is related to the lower *VOP* of the \*PME since the ability to more consistently portray truth results in better discrimination.

Another possible advantage of PME is that its overdispersion may somewhat alleviate the negative effects of undersampling. Recall that undersampling results in an overforecast of high *FP* and an underforecast of low *FP* (see I.B.4 and II.C.3). With an excessive ensemble spread, the PME produces slightly reduced high *FP* values and slightly lower low *FP* values, thus reversing the undersampling effect. While this may just sound like a statistical trick, it is an actual benefit. A small ensemble with slightly excessive spread does a better job at representing the PDF tails compared to an ideal ensemble with the same number of members. It may actually be advantageous for a small ensemble to be slightly overdispersive.

## 2. Mesoscale: ACME<sup>core</sup> vs. ACME<sup>core+</sup>

Thus far, we have concluded that the perturbed-model approach of ACME<sup>core+</sup> fails to capture all the model uncertainty. The question that now remains is, putting the limitations of ACME<sup>core+</sup> aside, did inclusion of model diversity improve SREF on the mesoscale, and if so, how? This section presents a detailed comparison of ACME<sup>core</sup> vs. ACME<sup>core+</sup>, revealing that inclusion of model diversity is critical for a mesoscale SREF.

## a) Dispersion

We return to the dispersion diagrams (Figure 47) to begin comparing the two mesoscale SREF systems. In Figure 47a & b, \*ACME<sup>core+</sup> made only a small improvement to the severe underdispersive quality of \*ACME<sup>core</sup> for synoptic-scale parameters. However, \*ACME<sup>core+</sup> made more of an improvement for *MSLP* than for *Z*<sub>500</sub> because, as a surface parameter, *MSLP* is more influenced by model error. The influence of model error is even more pronounced in Figure 47c & d in which the dispersion of surface, mesoscale parameters is examined. The poor dispersion of ACME<sup>core</sup> is more evident (especially for *T*<sub>2</sub>) and the improvement by ACME<sup>core+</sup> is more significant. Also notice that the *WS*<sub>10</sub> and *T*<sub>2</sub> *MSE* results are much closer to saturation (i.e., reaching  $\sigma_c^2$ ), but still well below. For *MSLP* and *Z*<sub>500</sub>, the *MSE* results at the 48-h lead time were only about 5% of the way toward saturation, but are about 30% of the way for *WS*<sub>10</sub> and *T*<sub>2</sub>, thus confirming the lower predictability of the mesoscale parameters.

Focusing on Figure 47d, there is a pronounced diurnal signal in the *T*<sub>2</sub> *MSE* but very little error growth. The lack of error growth is not because of error saturation but because *T*<sub>2</sub> is primarily locally forced rather than synoptically forced. The variability in *T*<sub>2</sub> is determined mostly by the diurnal heating and only secondly by the large-scale flow in which errors grow. The bulk of the error is therefore determined by the model's deficiencies. Errors increase during the night to a maximum right before sunrise, then reach a minimum by midday—likely due to the difficulty in modeling the planetary boundary layer (PBL). The model can not accurately describe the collapse of the PBL and formation of inversions at night so *T*<sub>2</sub> is often greatly in error then. During the day, the model may not get the PBL quite right, but low level mixing is normally present to some degree so *T*<sub>2</sub> errors are not as extreme. Notice that the diurnal signal of the *MSE* correlates very strongly with the \*ACME<sup>core+</sup> spread but only weakly with the

\*ACME<sup>core</sup> spread—another indication that \*ACME<sup>core+</sup> is a significantly better system for representing the forecast uncertainty.

Focusing on Figure 47c, there is a slight diurnal signal in the  $WS_{10}$  MSE and significant error growth. The  $WS_{10}$  error growth is dramatically higher than that of  $T_2$  because, even though  $WS_{10}$  is a locally forced mesoscale parameter, it also is strongly influenced by the synoptic flow since surface winds are forced by surface pressure, which is dependent upon the deep atmosphere. The resulting error growth from the ICs alone allows the  $WS_{10}$  spread to be much closer to reaching the target variance compared to  $T_2$ . However, the large increase in spread by \*ACME<sup>core+</sup> shows that inclusion of model diversity is still very important for a complete representation of uncertainty in  $WS_{10}$ .

\*ACME<sup>core+</sup> gives some MSE improvement over \*ACME<sup>core</sup> for  $WS_{10}$  and notable improvement for  $T_2$ , which appears contrary to the Chapter II statement that the ACME<sup>core+</sup> members are generally inferior to the members of ACME<sup>core</sup>. Evidently, the inferiority is mainly in systematic error for, once the bias is removed, the EF mean of \*ACME<sup>core+</sup> is superior. So not only did ACME<sup>core+</sup> increase EF spread toward the target variance, it also narrowed the gap further by lowering the target variance (i.e., reducing the MSE of the EF mean).

In Figure 47a & b there is an estimate of the IC differences since  $MSLP$  and  $Z_{500}$  exist as IC fields. For Figure 47c & d, there is no initialization of  $WS_{10}$  and  $T_2$  so their initial uncertainty can not be considered. We can however state that the uncertainty in all parameters from which  $WS_{10}$  and  $T_2$  are derived were exactly the same at initialization in ACME<sup>core</sup> and ACME<sup>core+</sup> since both systems used the same ICs. Most of the increased spread provided by \*ACME<sup>core+</sup> is likely realized in the very first forward time step of the model (the first 36 s of the forecast period) in which the different model options and SBPs produce much different values of derived surface variables. During the rest of the forecast period the gap between the \*ACME<sup>core</sup> and \*ACME<sup>core+</sup>

spread increases only slightly as some of the differences provided by model diversity project onto growing modes. For a parameter with a strong synoptic influence like  $WS_{10}$ , almost all of the predictability error growth (revealed by the EF spread of  $*ACME^{core}$ ) comes from uncertainty in the ICs, and the additional spread from model diversity (revealed by the EF spread of  $*ACME^{core+}$ ) simply adds a constant correction toward statistical consistency. Early in the forecast period the additional spread provided by  $*ACME^{core+}$  is a larger fraction of the total spread, and is thus more important to include in the SREF.

Compared to a synoptically-influenced parameter like  $WS_{10}$ , including model diversity for a parameter like  $T_2$  is even more important since the vast majority of the forecast error is due to the model uncertainty and not the IC uncertainty—an unusual finding discussed further below. In Figure 47d, the low spread of  $*ACME^{core}$  shows how little of the forecast error originates from the ICs and confirms the lack of error growth. The fact that  $*ACME^{core+}$  spread is still far below what is required for statistical consistency indicates that much more model diversity is required.

The VRHs of Figure 48 are ordered from the parameter in which inclusion of model diversity makes the least difference ( $Z_{500}$ ) to the parameter where it makes the most difference ( $T_2$ ).  $*ACME^{core+}$  provided only minor improvement for the synoptic-scale parameters ( $Z_{500}$  and  $MSLP$ ) in which the forecast error is dominated by error growth from the ICs. For the more mesoscale parameters of  $WS_{10}$  and  $T_2$ , it is evident that  $*ACME^{core}$  is extremely poor at portraying truth. With the added spread of  $*ACME^{core+}$ , the VRHs are adjusted toward uniformity and the large  $VOP$  is cut in half. However, as we saw with the dispersion diagrams, it is also obvious that  $*ACME^{core+}$  is still far from being statistically consistent.  $*ACME^{core+}$  fails to represent a considerable amount of the uncertainty that is present and truth is still not portrayed far too often.

From Figure 47 and Figure 48, we conclude that the closer to the surface and smaller in scale a phenomenon, the more difficult it is to represent its uncertainty and the more model uncertainty



appears to play a part. We also conclude that inclusion of model diversity in a mesoscale SREF is critically important for complete representation of forecast uncertainty and that the relative role of IC and model uncertainty depends upon the parameter as well as the weather regime.

A question that remains is what is the reason for the low dispersion of \*ACME<sup>core+</sup>? From the previous section on comparing the \*PME and \*ACME<sup>core+</sup> we concluded that the use of a LAM limits dispersion of \*ACME<sup>core+</sup> only marginally and the major difference between the two systems is their relative amount of model diversity. That would suggest we need to increase model diversity of \*ACME<sup>core+</sup> by expanding the perturbed-model method through more and/or larger perturbations. However, there is another possible source of the low dispersion problem of \*ACME<sup>core+</sup>: limitations imposed by a finite model grid resolution.

Smagorinsky (1969) demonstrated that increasing model resolution increases dispersion of the model since higher resolution can represent additional scales of motion. For an ensemble, differences among the members can only exist at the scales represented within the model, so there can be no difference (i.e., no dispersion) at unrepresented scales. Part of the low dispersion of ACME<sup>core+</sup> is likely due to the limited capability of the 12-km members to reveal different possibilities at small scales. Increasing model resolution should generate more useful spread among the members by capturing more diversity in smaller scale motions. To test this hypothesis with our research data, we can directly compare the dispersion over matching grid points on the 12-km and the 36-km domains (i.e., compare every third point in the 12-km domain to the subsection of points in the 36-km domain that overlays the 12-km domain). Results for  $WS_{10}$  from \*ACME<sup>core+</sup> reveal that the ensemble spread on the 12-km domain is an average 27% higher than on the 36-km domain (Figure 49b). There is likely an asymptotic limit to how much more dispersion can be produced by finer model resolution, but we suspect that significantly higher dispersion could be realized by increasing model resolution to a few km.

## b) Skill and Utility

Just as  $*ACME^{core+}$  did not significantly improve the severe underdispersion of  $*ACME^{core}$  for *MSLP*, Figure 43 shows that the  $*ACME^{core+}$  *BSS* for an *MSLP* event is about the same as for  $*ACME^{core}$ . For  $T_2$  on the other hand, Figure 44 shows that  $*ACME^{core+}$  is more skillful, which means that there is value in the additional spread provide by  $*ACME^{core+}$ . The drastically different *BSS* improvements for *MSLP* and  $T_2$  can be partly explained by the fact that  $T_2$  error is highly determined by the model, which makes inclusion of model diversity more important.

The successful *BSS* results for  $T_2$  appear to contradict the requirement of Murphy (1988) and Palmer et al. (1990) that for an EF to have a chance at being effective, the portion of forecast error due to IC uncertainty must be larger than the portion due to model uncertainty. It is clear from Figure 47 that model error dominates  $T_2$  forecast error, and it is also clear from Figure 44 that skillful *FP* was produced for a  $T_2$  event.  $T_2$  is however not a state variable but a derived variable of the PBL scheme. The skill of  $T_2$  *FP* depends on many other variables for which IC uncertainty may be larger. The requirement that IC uncertainty be larger than model uncertainty applies to the forecast as a whole (over all dimensions) and not a limited slice of phase space.

Besides that fact that *MSLP* is mainly a synoptically forced parameter, another factor for the weak  $*ACME^{core+}$  *MSLP* skill improvement is that model error plays a much greater role over land than over water. Convection, wind flow over complex terrain, variations in radiative effects, etc. all require more detailed parameterizations and schemes within the model and thus more opportunity for model error. To demonstrate the increased need for model diversity over land, we recomputed *BSS* using ocean-masked data (i.e., use only grid points over land). For an equitable comparison of *MSLP* *BSS*, we raised the event threshold to keep a similar *SC* (and similar *unc*) as that in Figure 43 (i.e., *MSLP* climatologic PDF is shifted upward over land). In Figure 51, the improvement in *BSS* by  $*ACME^{core+}$  over  $*ACME^{core}$  is about 2 h whereas it was not measurable

for the full domain. Furthermore the improvement by \*PME is up from 11 h to roughly 18 h. The increased improvements by the SREF systems with model diversity suggest that including model diversity over land is more important.

Figure 52 shows a similar result for ocean-masked  $T_2$  on the inner domain. Note that we did not alter the event threshold here since *unc* of the event was only slightly higher compared to the full domain results in Figure 44, which of course makes ocean-masked *BSS* relatively lower. The difference in ocean-masked and full-domain  $T_2$  *BSS* improvement by \*ACME<sup>core+</sup> is most evident by the plots of Figure 53a & b. The afternoon (18Z – 24Z and 42Z – 48Z) dip in improvement in both plots is associated with a relatively low increase in *res* by \*ACME<sup>core+</sup> during these periods. The contribution to *BSS* improvement by *rel* and *res* are comparable except during the afternoon when the *rel* improvement decreases somewhat and the *res* improvement becomes minimal. (Notice however that there is still a large improvement by bias removal in the afternoon.) A possible explanation for the low afternoon *res* improvement is that as previously noted, the variability of  $T_2$  is relatively lower compared to the nighttime  $T_2$ . While including model diversity did increase EF spread in the afternoon (i.e., Figure 47d), the gain was not as spectacular as at night in which a much greater widening of the forecast PDF was required.

Figure 54 shows the *BSS* for ocean-masked  $WS_{10} > 18$  kt (an operationally significant event) and supports the conclusion that \*ACME<sup>core+</sup> provides the best *FP* and that bias correction and inclusion of model diversity in a SREF is critical. However, it is also evident in Figure 53c & d that the improvement by \*ACME<sup>core+</sup> was much less for  $WS_{10}$  compared to the improvement for  $T_2$  because the increase in spread by model diversity did not make as significant an impact as with  $T_2$ , evidenced by Figure 48c & d. Another observation in Figure 54 is that the difference between ACME<sup>core</sup> and ACME<sup>core+</sup> (either before or after bias correction) is greater earlier in the forecast cycle, suggesting that including model diversity is more important for earlier lead times. This

supports the  $WS_{10}$  dispersion diagrams showing that model error contributes a larger fraction of the total forecast error in the earlier lead times.

The  $WS_{10}$  results demonstrate that the greater improvement over land when including model diversity is not simply a statistical artifact. That is, one might argue that for  $T_2$  the higher *unc* and lower *BSS* of the ocean-masked data provide more of an opportunity for improvement so the comparison is unfair. The counter to that argument is that for  $WS_{10}$  there is lower *unc* and higher *BSS* (not shown) but the same result of greater improvement over land.

Figure 55 gives the relative operating characteristic skill score (*ROCSS*) results for the three events we studied. In general *ROCSS* provides a more obvious analysis of the utility of FP and is considered to be an upper bound of overall forecast value whereas the *BSS* is the lower bound (Jolliffe and Stephenson, 2003). (I.e., comparing Figure 55 to Figure 51, Figure 52, and Figure 54, the *ROCSS* is consistently higher than the *BSS*.) Figure 55 confirms the higher utility of  $*ACME^{core+}$  over  $*ACME^{core}$ . This analysis also shows that the bias removal may not have worked well for  $WS_{10}$  or for *MSLP* in the late afternoon. However, this is not a conclusive result since as Marzban (2003) pointed out, the area under the ROC is not good at discriminating between two EF systems that performed well for a certain event.

### C. ACME and Analysis Uncertainty

Recall that the purpose of the ACME with its additional members was to mitigate the problems associated with a small ensemble by further sampling analysis uncertainty, thereby producing more ICs and boosting ensemble size. This was accomplished by mirroring each of the core analyses about the centroid analysis. In this section we will show that ACME was successful at generating new ICs (based on the core analyses) that produced valid, unique forecasts with valuable information. However, ACME failed to significantly improve overall skill commensurate with the increase in ensemble size.

## 1. Skill and Utility

Figure 56 is a comparison of the overall deterministic performance of the ACME members. The core members are named after the analysis that provided their IC and LBCs (listed in Table 3). The centroid forecast is called 'cent', and, following Equation (19), a mirrored member's name is the name of its source analysis primed. (E.g., the *cmcg'* member was run using the IC and LBCs created by mirroring the *cmcg* across the centroid analysis.) Figure 56 shows that the mirrored members are basically on a par with the core members. If the mirrored members were not valid forecasts, their average *RMSEs* and rankings would stand out as higher than the core's.

We did not include a similar plot for  $T_2$  in Figure 56 since, as discussed above,  $T_2$  skill is so heavily dependent on the model that varying the IC makes little difference (recall Figure 31). For mesoscale parameters with error that is primarily model dependent, expanding or improving ICs contributes almost nothing to improving SREF performance. To improve *FP* of  $T_2$ , one should concentrate on representation of the model's stochastic error since, in general, any reasonable set of ICs may be used. There is little need for an approach such as ACME for these types of parameters.

A very positive result from the ACME system is the excellent performance of *cent*. It has an average *RMSE* equal to or better than the best core member and has the best average ranking. The high skill of *cent* is what convinced us to use the centroid analysis as verification. The centroid analysis is normally the best estimate of truth since averaging of the eight analyses likely cancels out a large portion of the errors that exist in the individual analyses (Richardson, 2001a). It may be argued that the superior performance of *cent* is a statistical artifact of its smoothness at the initialization. However, *cent* is not a smoothed average of other forecasts but a complete MM5 run containing information on all scales. Note that the centroid analysis includes the

obviously inferior analysis data from tcwb, which likely degraded cent. The superiority of cent would likely stand out even further if tcwb was omitted from the centroid analysis.

The fact that tcwb stands out as an inferior member again raises the question of when does a member add benefit to an ensemble. In section II.B.1 (discussion of the perturbed-model method) we explained how a member with lower average skill can still add value to the ensemble if it can occasionally perform better, but there is logically a limit to that effect. If a member rarely or never performs well, it may in fact degrade the overall performance of the ensemble. To test if that is the case with tcwb, we removed tcwb from both \*PME and \*ACME<sup>core</sup> and computed *BSS* for  $P(MSLP < 1001 \text{ mb})$  to compare with the full ensembles' *BSS* results. Note that there should be a slight reduction in *BSS* (on the order of 0.1%) due to the decrease in ensemble size from 8 to 7 members. Figure 57 shows that the 7-member ensembles with tcwb withheld performed better, indicating that tcwb is indeed harmful to our SREF systems. (The effect shows up more clearly in \*PME than in \*ACME<sup>core</sup> since the \*ACME<sup>core</sup> members are much more similar.) It may be that tcwb can occasionally perform well but evidently it is inferior so much of the time that its overall effect is to degrade the estimation of the forecast PDF. As a check on this effect, we also found *BSS* after removing a superior member (ukmo) and after removing a member with average skill (ngps). Figure 57 shows that without ukmo, probabilistic skill was significantly reduced and without ngps, was reduced 30-50% as much as the reduction from withholding ukmo. We conclude then that the higher the deterministic skill of a member, the more value it adds to an ensemble. Lower skilled members can add value to an ensemble but must perform well a significant portion of the time.

A notable difference between the mirrored and the core members is that the mirrored members do not have any outstandingly good members whereas the core has the avn and ukmo. Furthermore the average performance of the core members is itself mirrored in the mirrored

members. (E.g., ukmo performs great but ukmo' performs poorly; gasp performs poorly but gasp' performs well). It is easy to see how combining information from the good cent IC with the bad tcwb IC would result in the tcwb' forecast being better than the tcwb forecast. However, there were cases where the tcwb' forecast would also outperform many ACME members including cent. Evidently the vector between tcwb and cent can occasionally be a very good estimate of the analysis error. Unfortunately, we found no clear way to identify such cases a priori.

If we were to accept that the ACME members are additional samples from the same forecast PDF as  $ACME^{core}$  and that  $ACME^{core}$  exhibited statistical consistency (which of course it does not), then we would expect ACME to improve  $BSS \sim 0.03$  over  $ACME^{core}$  due solely to the increase in ensemble size from 8 to 17 (see section II.C.3). Figure 58 and Table 8 show that \*ACME came nowhere close to this expectation and performed about the same as \* $ACME^{core}$ . Varying the event thresholds produced similar results (not shown). Of the three parameters studied, \*ACME was only able to slightly improve  $MSLP$  because, due to its synoptic nature,  $MSLP$  is more sensitive to IC variations.

An explanation for the lack of improvement by \*ACME is that producing more samples from a deficient forecast PDF may result in a more detailed, but still deficient, description of the possible future states. Recall that an ideal ensemble with a small  $n$  has greater variance in its sampling distributions of the mean and spread so that the  $n$  members are unable to consistently represent their PDF and  $FP$  skill is degraded. Increasing  $n$  for an ideal ensemble allows the forecast PDF to match the true PDF more consistently. ACME does result in a more consistent representation of the PDF from which the members are drawn, but, since it is not an ideal ensemble, ACME does not provide a better representation of the true PDF. In other words, ACME suffers from the same basic problems as  $ACME^{core}$  but ACME does improve upon the poor self-consistency of  $ACME^{core}$ . The original hope of ACME was that it could not only

provide new valid samples but also expand the forecast PDF to sample regions of phase space where  $ACME^{core}$  failed to portray truth. In the next section we show evidence that ACME was able to encompass more truth but did little toward better portrayal of truth.

## 2. Dispersion

The dispersion diagram for *MSLP* (Figure 59) shows that the average spread of \*ACME is slightly lower compared to the spread of  $ACME^{core}$ . The lower spread may be due, in part, to the MM5 preprocessing of the ICs. Recall that the mirroring perturbation factor was designed so that ACME would have the same initial spread as  $ACME^{core}$ . However, the MM5 preprocessing adjusts the fields to obtain vertical balance, thus reducing the initial variance. Additionally, Figure 59 shows that on average, errors grow more slowly in \*ACME compared to  $ACME^{core}$  so perhaps mirrored perturbations sometimes lie off the model attractor and must reconverge before growing. Figure 59 shows that \*ACME has roughly the same (perhaps slightly worse) lack of statistical consistency as  $ACME^{core}$ , supporting the conclusion that the additional sampling of ACME did not result in an improved representation of the forecast PDF.

This disappointing conclusion is tempered by the VRH results. The lower missing rate (*MR*) and *VOP* of \*ACME in Figure 60a – c indicate that \*ACME was able to produce many verification values that were not represented in  $ACME^{core}$ . However, the missing rate error (*MRE*, difference from the ideal  $MR$  of  $2 / (n+1)$ ) reveals that \*ACME performed about the same or slightly worse than  $ACME^{core}$  since the VRHs have about the same degree of nonuniformity. In an absolute sense, \*ACME encompassed more truth than  $ACME^{core}$  and provided valuable information. However, considering that the expected amount of encompassed truth depends upon ensemble size, \*ACME performed roughly the same or worse than  $ACME^{core}$ .

The VRH comparison for  $T_2$  in Figure 60c is quite different than that for synoptically forced  $Z_{500}$  and *MSLP* Figure 60a & b. As discussed above, when the forecast error in a parameter is



dominated by the model, additional variations in the ICs provide little benefit. Notice that the \*ACME  $T_2$   $MR$  decreased only slightly, the  $MR$  error greatly increased, and the  $VOP$  remained the same showing that \*ACME was not able to provide new information.  $WS_{10}$  in Figure 60c is an even mix of synoptic and model-driven error so \*ACME was able to add some new information.

The additional encompassing of truth by \*ACME for  $MSLP$  seems contradictory to the dispersion diagram results (Figure 59) that show a lower spread for \*ACME, but it is possible to decrease the domain-averaged standard deviation and still increase spread over limited areas where truth was previously not encompassed. Consider Figure 61 in which we compare occurrences of the verification in the extreme ranks and  $V_Z$  of \*ACME and \*ACME<sup>core</sup>. \*ACME greatly reduced the  $MR$  from \*ACME<sup>core</sup>, and almost did as well as \*PME. However, while \*ACME helped reduce the really high  $V_Z$  values of \*ACME<sup>core</sup>, there was actually a slight  $VOP$  degradation by \*ACME compared to the much improved \*PME  $VOP$ . The conclusion is that \*ACME did very little toward portraying truth better. Most of the reduction of the  $MR$  by \*ACME occurred where \*ACME<sup>core</sup> had already portrayed but not encompassed truth (i.e., where the verification occurred in an outside rank with a  $V_Z < 3s$ ). In other words, \*ACME only encompassed more truth where it was easy to do so and the major deficiencies (i.e., where truth really got away) of \*ACME<sup>core</sup> still largely remain in \*ACME.

Our final conclusion is that the ACME method is a sound way to further sample from the PDF defined by the core analyses, but ACME can not correct the deficiencies of ACME<sup>core</sup>. The core analyses occasionally miss key information, such as missing a shortwave trough, and no amount of mirroring can produce what was missed. Mirroring can, however, use the information available in the core analyses to create new plausible ICs, resulting in forecasts that provide a more thorough sampling of the forecast PDF of ACME<sup>core</sup>. Particularly effective is the ability of

ACME to better sample the forecast PDF tails, which reduces the *MR*. ACME cannot, however, magically sample far outside of the forecast PDF of ACME<sup>core</sup> to make up for what ACME<sup>core</sup> cannot represent.

#### D. Future Research

The results presented in this chapter suggest several areas of future research, the first of which is to investigate methods to improve the deficient dispersion of a mesoscale SREF. A possible technique to boost the synoptic-scale dispersion of our mesoscale SREF is to periodically nudge the MM5 forecast of each ACME<sup>core+</sup> member toward the large-scale model from which it was forced, thus imposing the beneficial large-scale dispersion of the PME onto the mesoscale SREF. In effect then, the PME would dictate the synoptic-scale error growth while the role of the ACME<sup>core+</sup> would be to show what that growth implies for mesoscale forecast uncertainty.

Nudging the ACME<sup>core</sup> members would likely improve SREF of mesoscale parameters that have a large component of synoptic forcing, such as  $WS_{10}$  and precipitation, but it would not improve statistical consistency for parameters such as  $T_2$  that are mostly model dependent. To improve the model-dependent parameters (and the others as well), additional ways to perturb the MM5 should be explored. This could mean simply further tuning the perturbations of ACME<sup>core+</sup>, such as increasing the magnitude of the SBPs, or digging deeper into MM5 to find additional model aspects to perturb.

Another approach to investigate for increasing dispersion of model dependent, mesoscale parameters is to increase model resolution. Such an increase would permit modeling of more scales of motion and should produce higher, more accurate dispersion among the ACME<sup>core+</sup> members. Higher resolution would also have the added benefit of reduced reliance on physical parameterizations so their errors would no longer have to be approximated. This is obviously the

most costly of solutions to the low dispersion problem so its benefit would have to be weighed against the processing requirements.

Another subject for future research is continued investigation into the multimodel vs. perturbed-model methods for representing model uncertainty. Nudging ACME<sup>core+</sup> members to eliminate the reduced dispersion of the LAM may allow a more fair comparison of the two methods. However, nudging would also have the potentially beneficial effect of driving the model diversity of the PME into ACME<sup>core+</sup> so that ACME<sup>core+</sup> would have both multimodel and perturbed-model components. Nudging would therefore only further cloud the distinction between the skill of PME and ACME<sup>core+</sup>. One way to truly compare multimodel and perturbed-model is to design an 8-member perturbed global model to compare against the PME. That is, make an ensemble using 8 perturbed-model versions of avn which use the PME ICs and compare skill of that ensemble to that of the PME.

Calibration of *FP* is a research issue that we mostly ignored except for the partial calibration by bias correction. The success of our bias correction raises some interesting questions concerning optimization of postprocessing. A calibration technique such as the weighted ranks method is designed to correct for systematic error of the ensemble as a whole and not by individual member. The advantage of a rigorous calibration is that it can correct for systematic dispersion problems besides model bias. However, a bias correction done on each member separately is much better at removing bias from the system since members may have much different biases. Therefore, the way to achieve the highest quality *FP* from a SREF system is likely to postprocess by bias-correcting each member followed by application of a calibration technique before producing *FP*. To test this one could compare the skill of bias-corrected only *FP*, calibrated only *FP*, and bias-corrected/calibrated *FP*.

A potentially valuable calibration technique is to take advantage of the fact that the ensemble members are not equally likely. We found that ACME<sup>core+</sup> members are not equally likely because of the varied skill levels of their ICs and choice of MM5 physics options. For EF in general, unequal skill among ensemble members is acceptable and even beneficial as long as each member can perform well some of the time. If the relative skill among the members could be determined a priori, it would benefit both probabilistic and deterministic forecasting. For FP, the relative skill levels could be translated into weights for use in calculating FP. For deterministic forecasting, the weights could be used to create a weighted ensemble mean as the best guess forecast.

The weights could be determined simply by the long-term average *RMSE* of the members. Unfortunately, Ebert (2001) showed that for a PME, weighting by long-term performance does not add value because the relative skill among members likely varies both spatially and temporally. To account for the temporal variation one could calculate the most recent forecasts' relative skills, which assumes there is a high level autocorrelation in a members' relative skill from one forecast cycle to the next. Simultaneously accounting for the spatial variation component is more difficult because if we are primarily concerned about a limited area (such as our inner 12-km domain) the relative skill may vary rapidly from one cycle to the next. A possible solution is to determine the relative skill among the members in the part of the atmosphere that will affect the 12-km domain in the current forecast cycle. This could be done using the MM5 adjoint model to define a 24-h sensitivity field for the low level flow over WA. The sensitivity field could then be multiplied by the *RMSE* field of the previous ACME<sup>core+</sup> run (using some representative parameter such as 700 mb GPH). This would reveal the relative ranking of the members based on how well they have recently represented the large-scale flow that will influence WA's weather in the current forecast cycle. That information is carried into

the current forecast cycle since the current cycle's analysis is largely based on the first-guess solution from the previous forecast cycle.

Lastly, the results of ACME<sup>core+</sup> showed that while there is definitely room for improvement, there is utility in SREF products. This fact needs to be further demonstrated to the weather forecast community through further studies and design of practical applications.

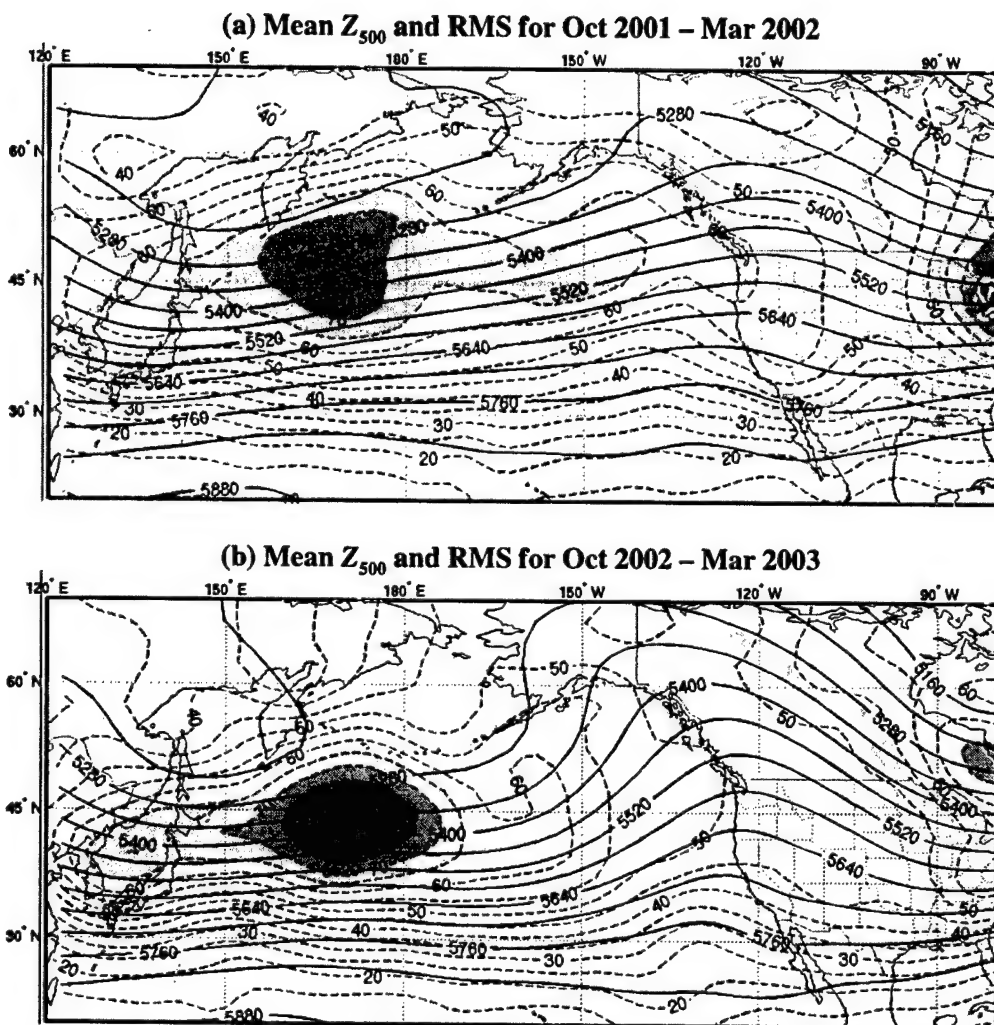


Figure 41. Mean  $Z_{500}$  and RMS of the time-filtered  $Z_{500}$  for the (a) 2001-2002 and the (b) 2002-2003 cool seasons (from Figure 5c & d in McMurdie and Mass, 2003).

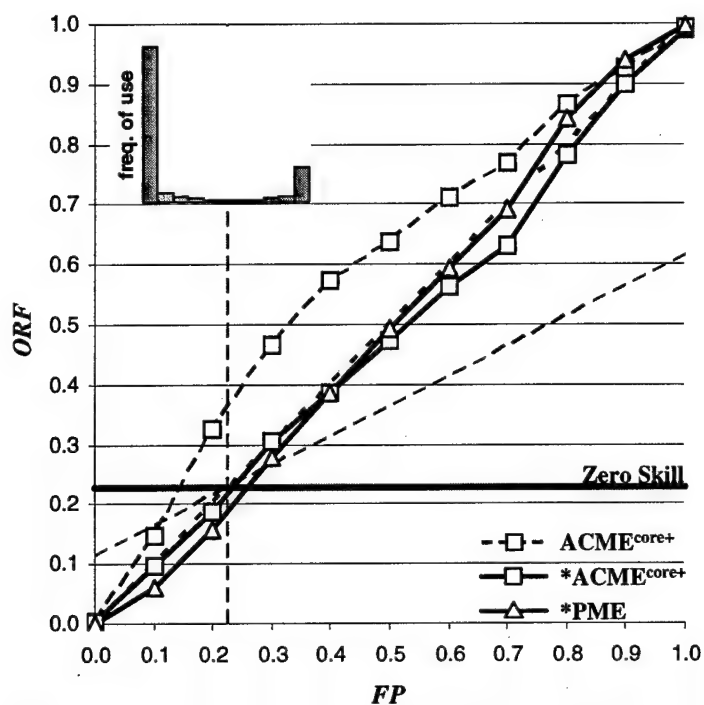


Figure 42. Reliability diagram for 36-h  $FP$  of the event  $MSLP < 1001.0$  mb. Compare  $ACME^{core+}$  and  $*ACME^{core+}$  for improvement by bias correction. Compare  $*PME$  and  $*ACME^{core+}$  for improvement of a multimodel system over a perturbed-model system. Data for this plot is found in Table 7.

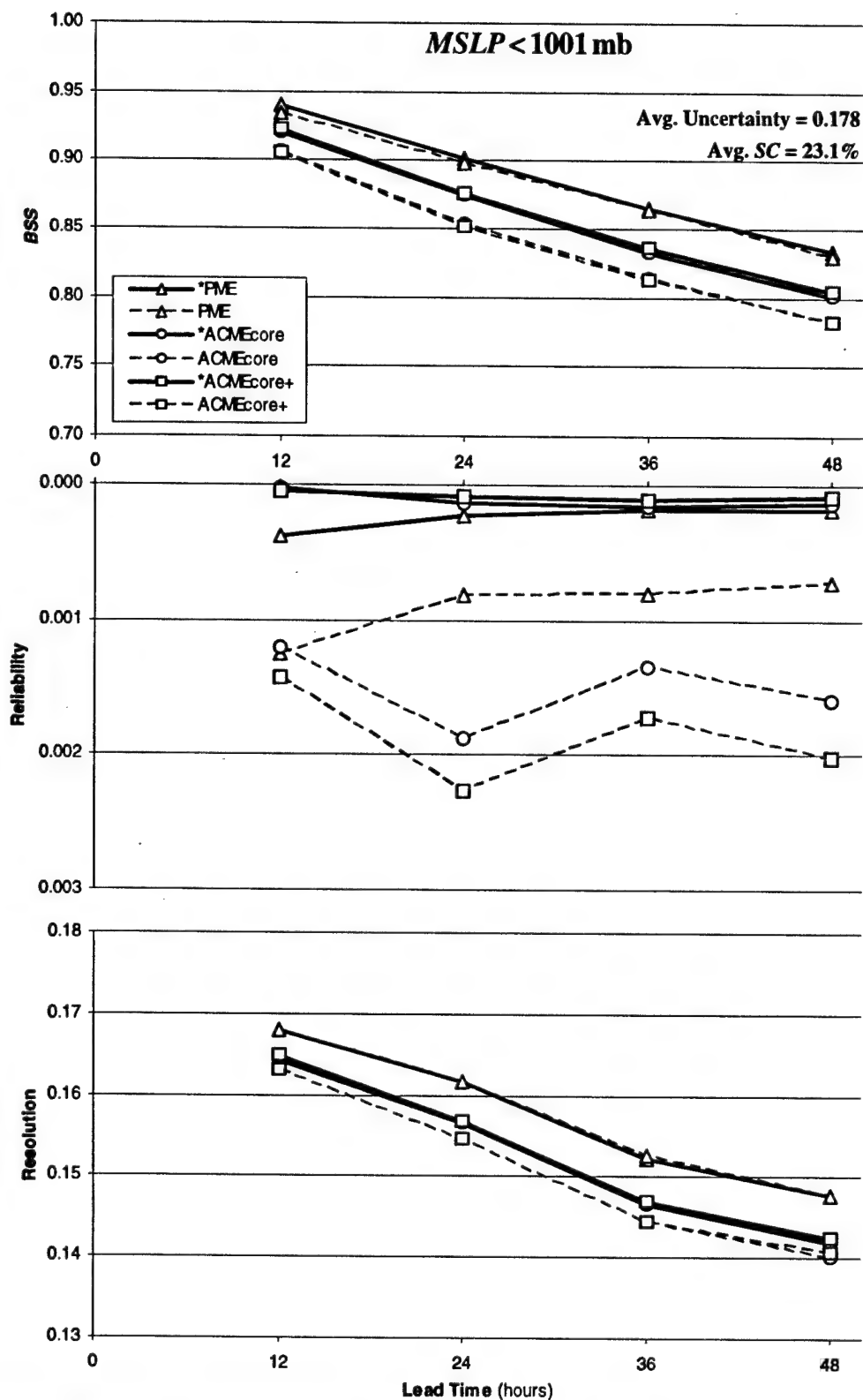


Figure 43. BSS and its components for FP of the event *MSLP < 1001.0 mb*.



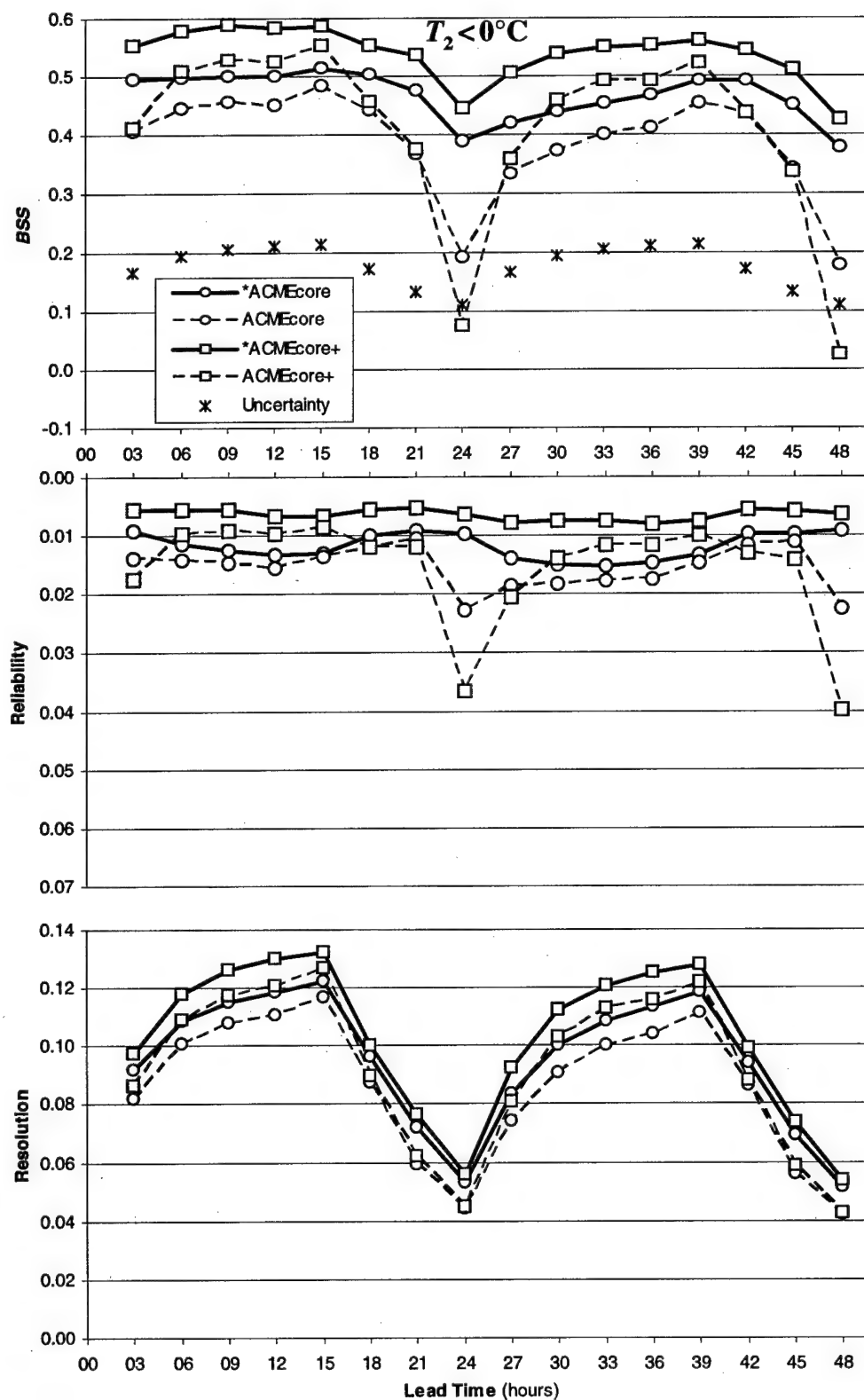


Figure 44. BSS and its components for FP of  $T_2 < 0^\circ\text{C}$ .

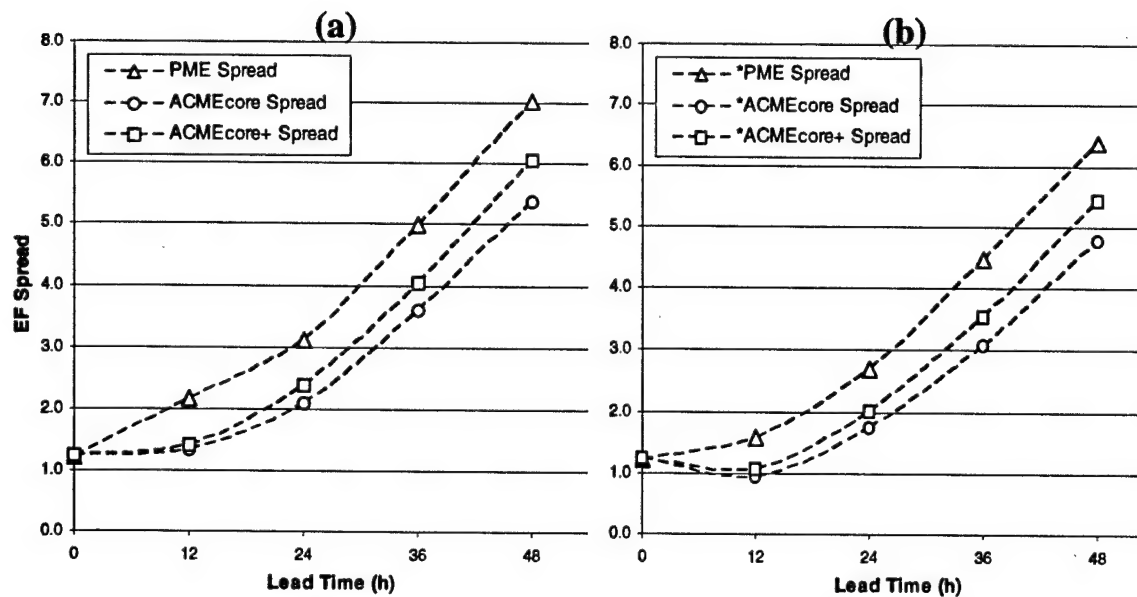


Figure 45. Dispersion diagram (without *MSE* of EF mean) for *MSLP* on the outer 36-km domain for (a) uncorrected forecasts, and (b) bias-corrected forecasts.

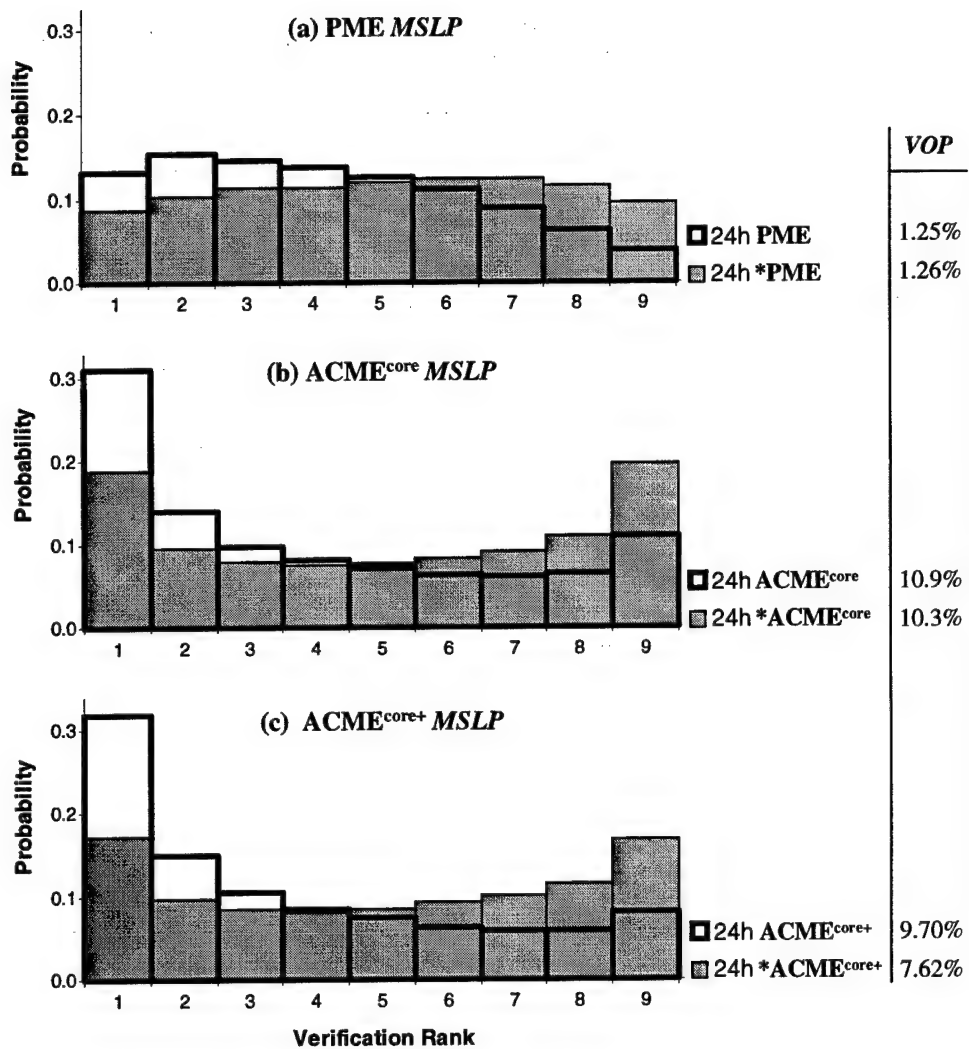


Figure 46. Effect of bias correction on VRHs of forecast *MSLP* with a 24-h lead time on (a) PME, (b) ACME<sup>core</sup>, and (c) ACME<sup>core+</sup>. The thick-lined VRHs were constructed from the original forecasts before bias correction and the shaded VRHs were constructed from the bias-corrected forecasts.

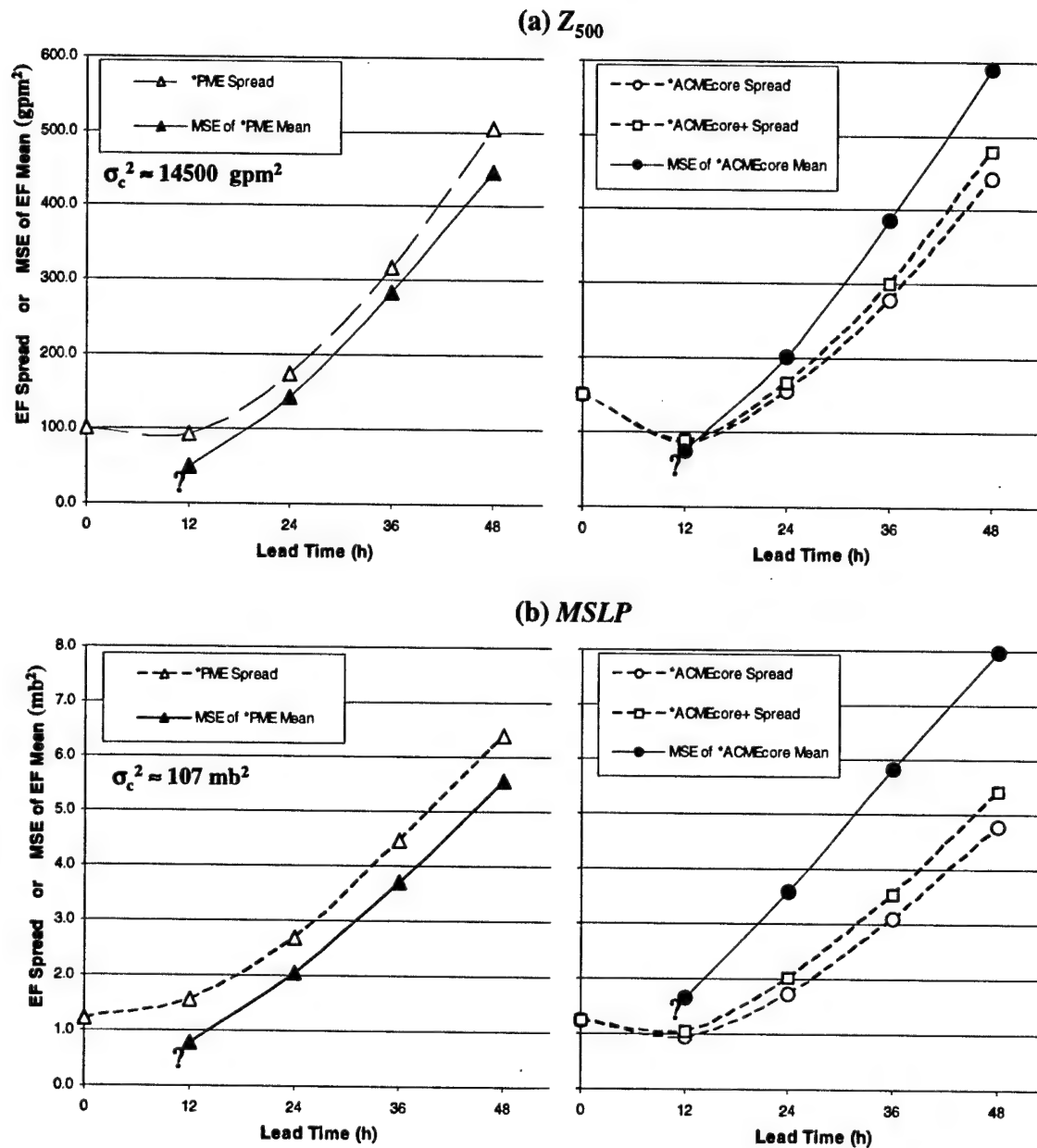


Figure 47. Dispersion diagrams showing EF spread (dashed lines with open points) and  $MSE$  of the ensemble mean (solid line with filled points) for bias-corrected forecasts of (a)  $Z_{500}$ , (b)  $MSLP$ , (c)  $WS_{10}$ , and (d)  $T_2$ . The PME results are shown separately from the  $ACME^{core}$  and  $ACME^{core+}$  since it is such a different system. In (a) and (b), only the  $MSE$  of the  $ACME^{core}$  mean is displayed since the  $MSE$  of the  $ACME^{core+}$  mean is  $<1\%$  different. Local times ( $L$ ) in 2 digit hours are marked on (c) and (d) to emphasize the diurnal signal of the error. The ? marks by the 12-h  $MSE$  results indicate that there is uncertainty in the result (see text).

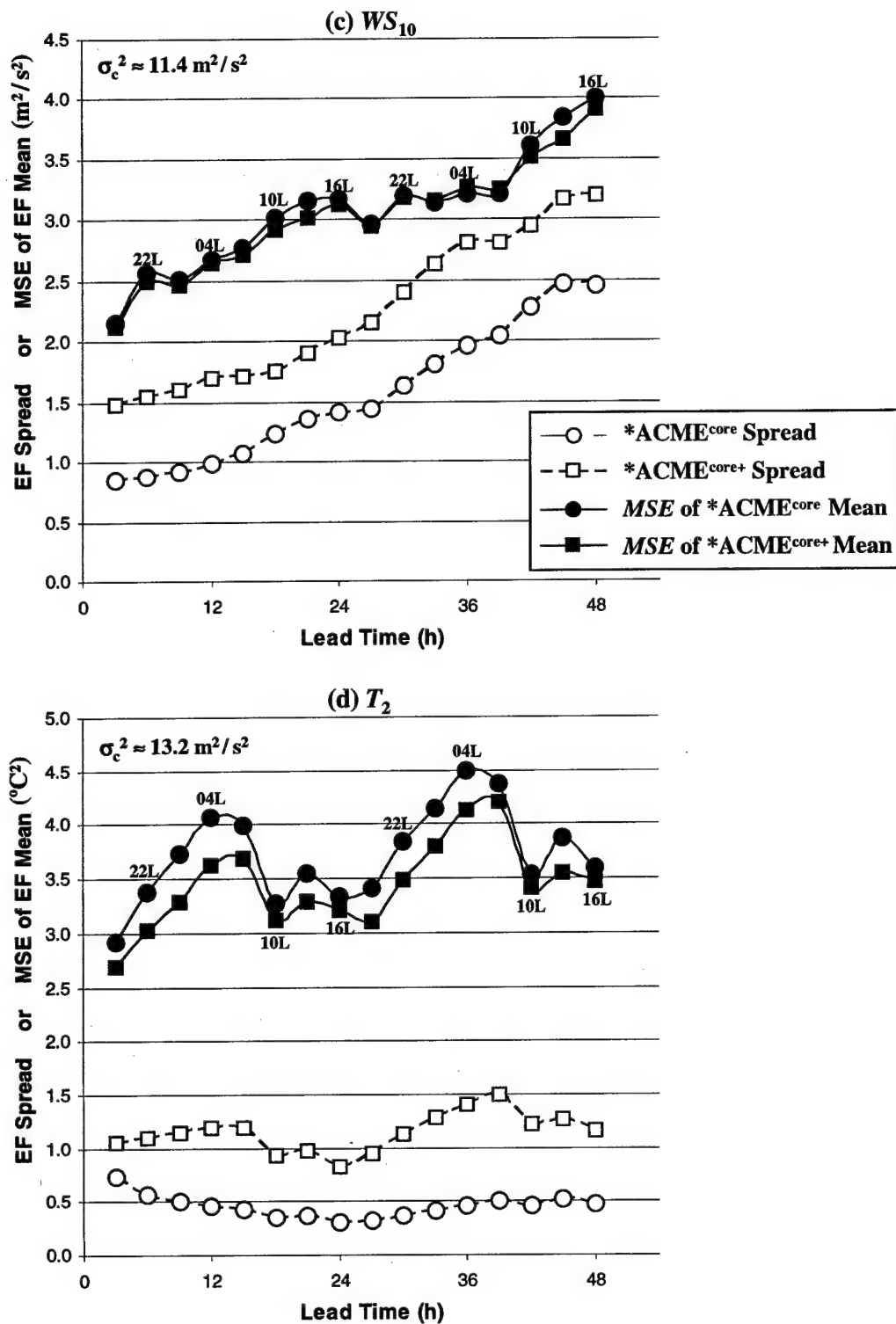


Figure 47 continued.

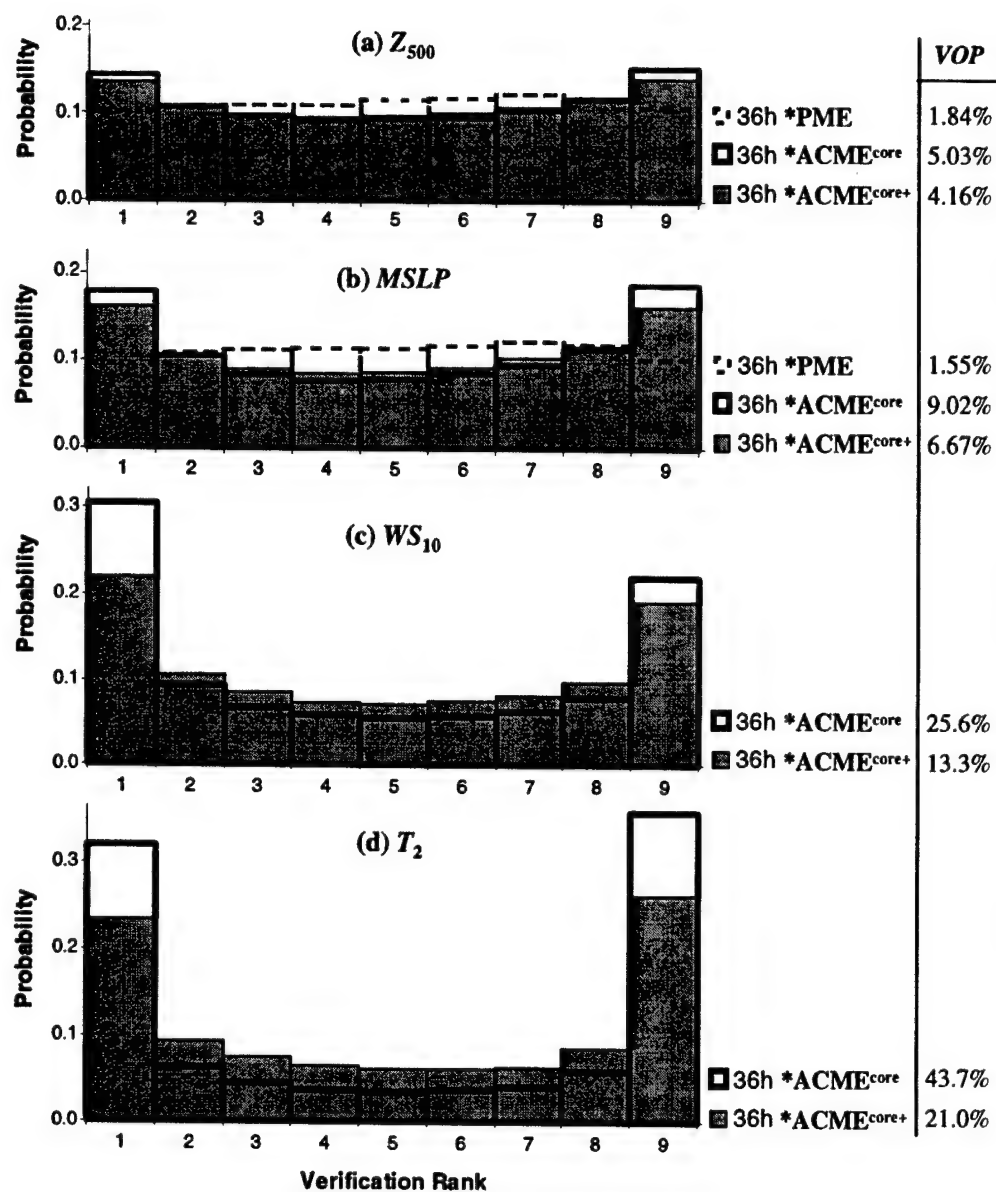


Figure 48. VRHs that show the impact of including model diversity for forecast (a)  $Z_{500}$ , (b)  $MSLP$ , (c)  $WS_{10}$ , and (d)  $T_2$  with a 36-h lead time. VRHs with a dashed outline were constructed from the \*PME forecasts. VRHs with a thick outline were constructed from the \*ACME<sup>core</sup> forecasts. Shaded VRHs were constructed from the \*ACME<sup>core+</sup> forecasts.

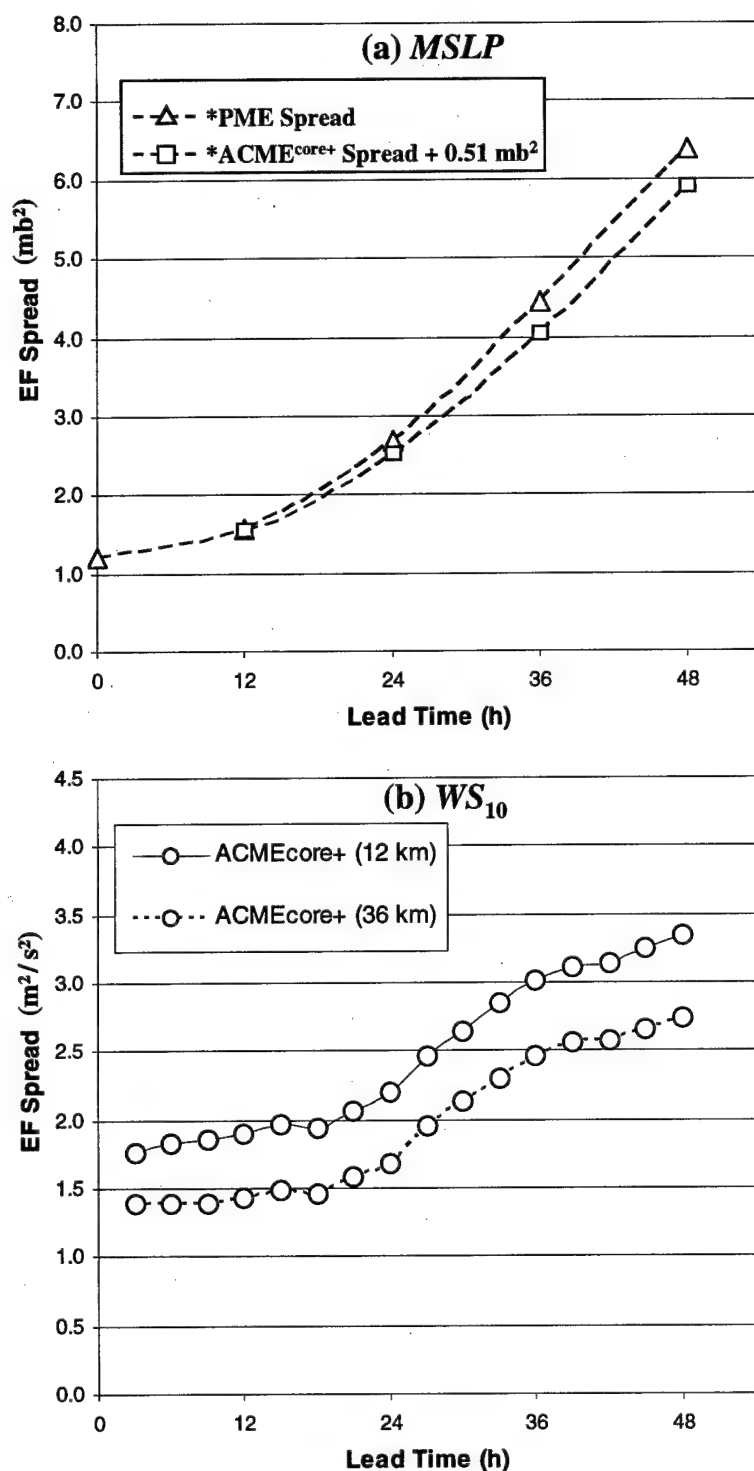


Figure 49. Special dispersion diagrams. (a) As in Figure 45b except that the difference in spread at 12 h between \*PME and \*ACME<sup>core+</sup> was uniformly added (at all lead times) to the \*ACME<sup>core+</sup> spread to make up for the spin-up effect in MM5. (b) Comparison of ACME<sup>core+</sup> spread for WS<sub>10</sub> from the 12-km and 36-km domains.

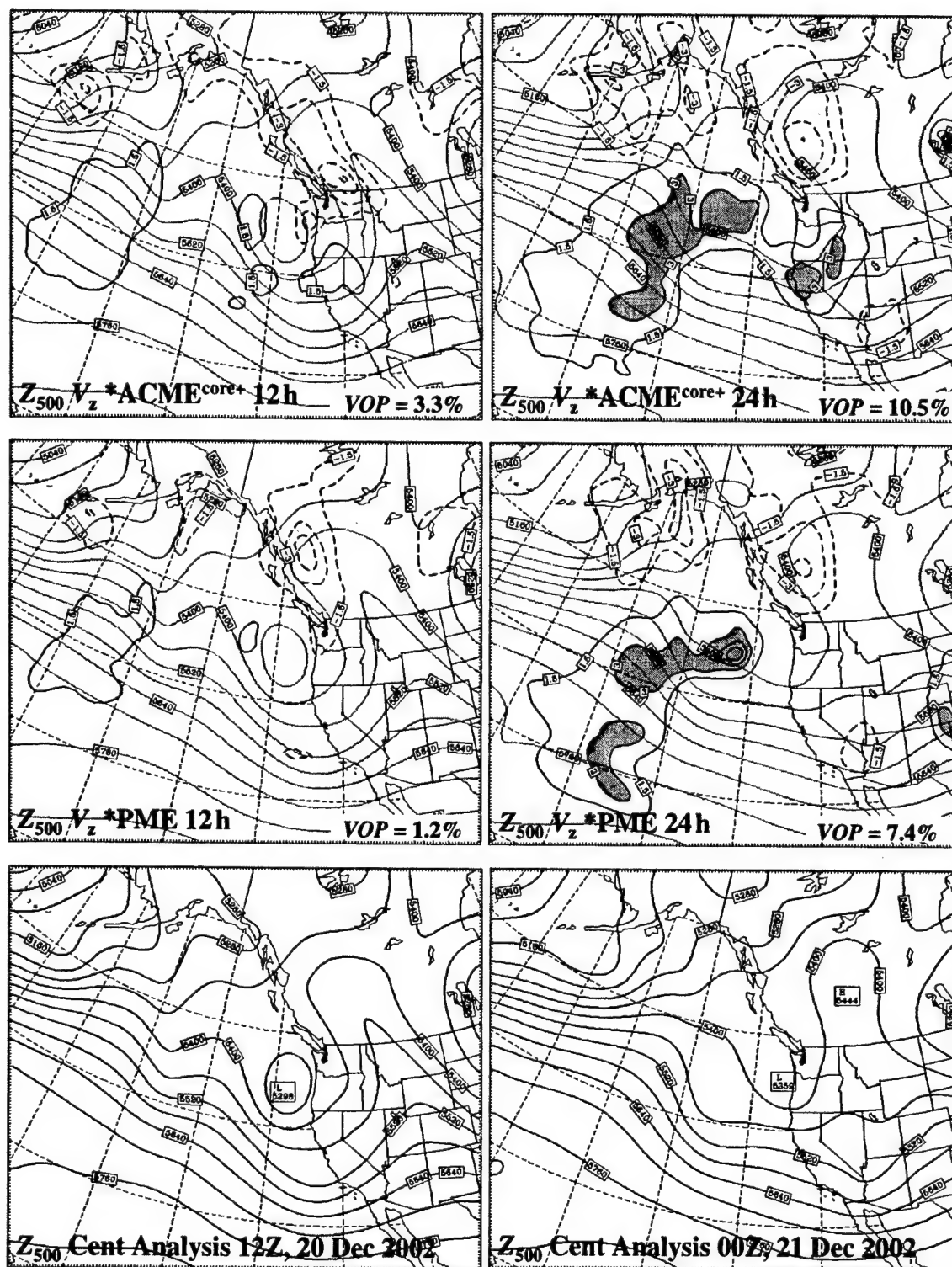


Figure 50.  $Z_{500}$  ensemble mean and  $V_z$  for \*ACME<sup>core+</sup> and \*PME forecast initialized 20 Dec 2002 at 00Z, along with corresponding centroid analyses. The inset VOP values are specific to this forecast case rather than averaged over all cases.  $V_z < -3$  is lightly shaded and  $V_z > 3$  is darkly shaded to show regions where the verification is an outlier with respect to the EF.



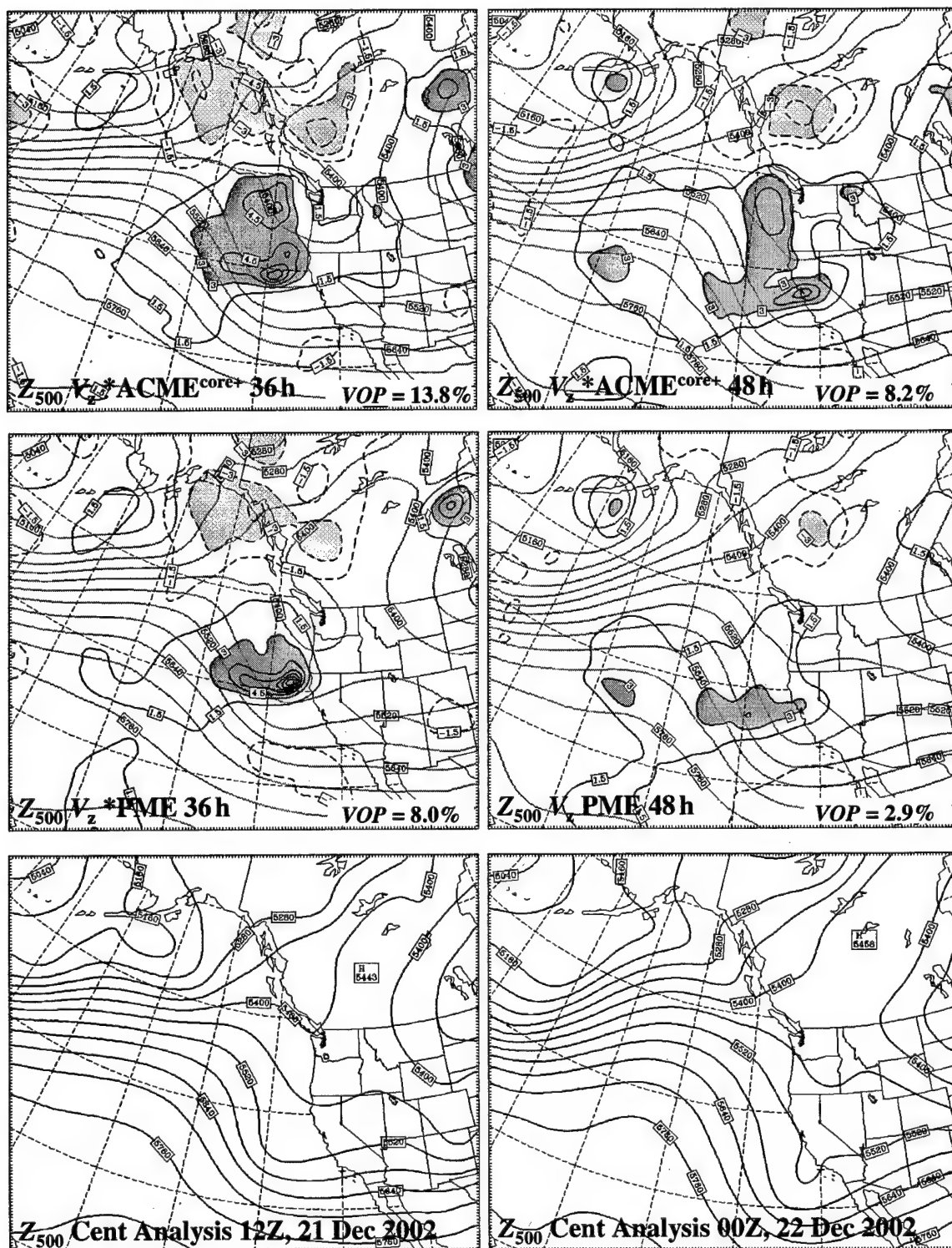


Figure 50 continued.

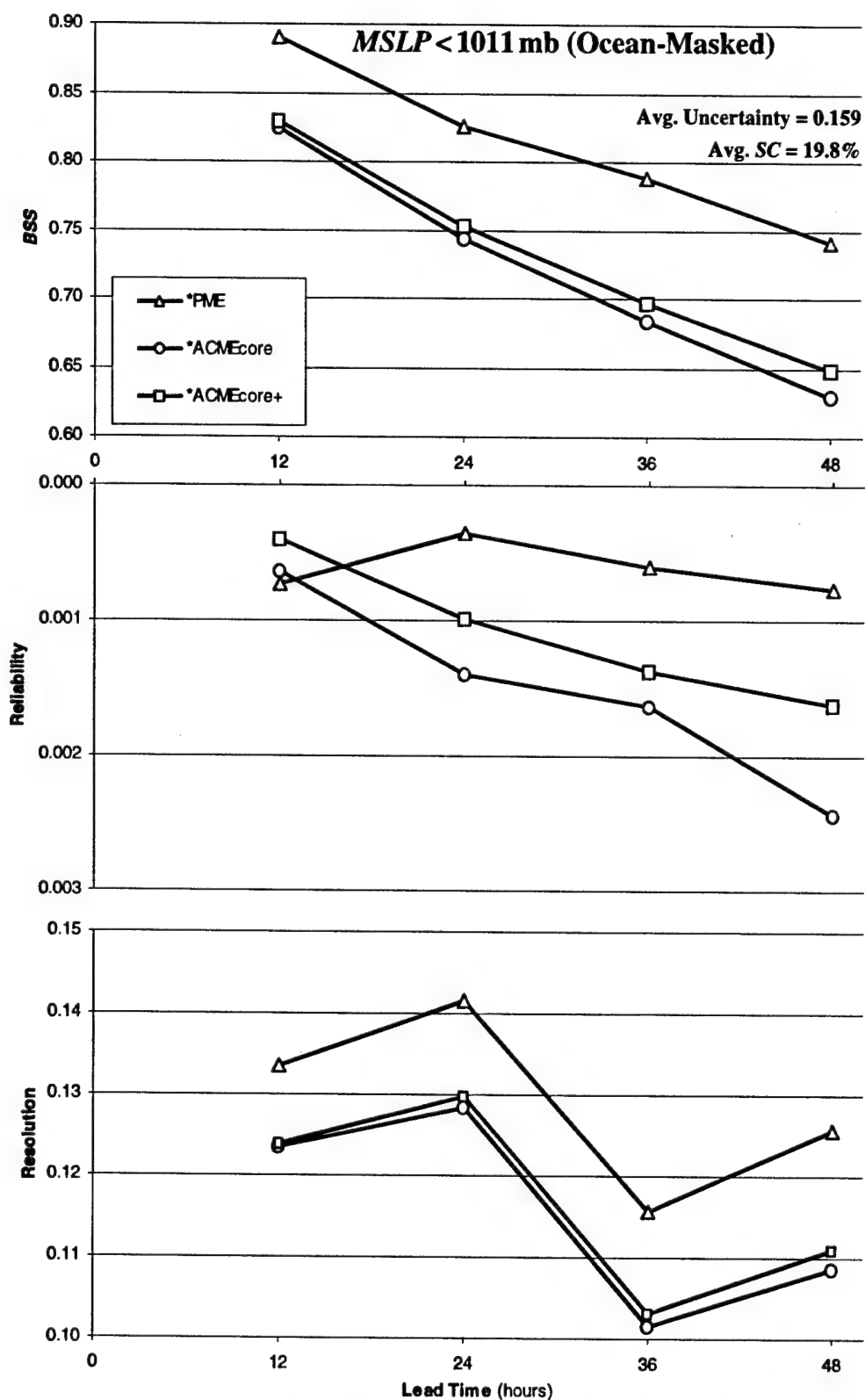


Figure 51. BSS and its components for FP of MSLP < 1011 mb using ocean-masked data.

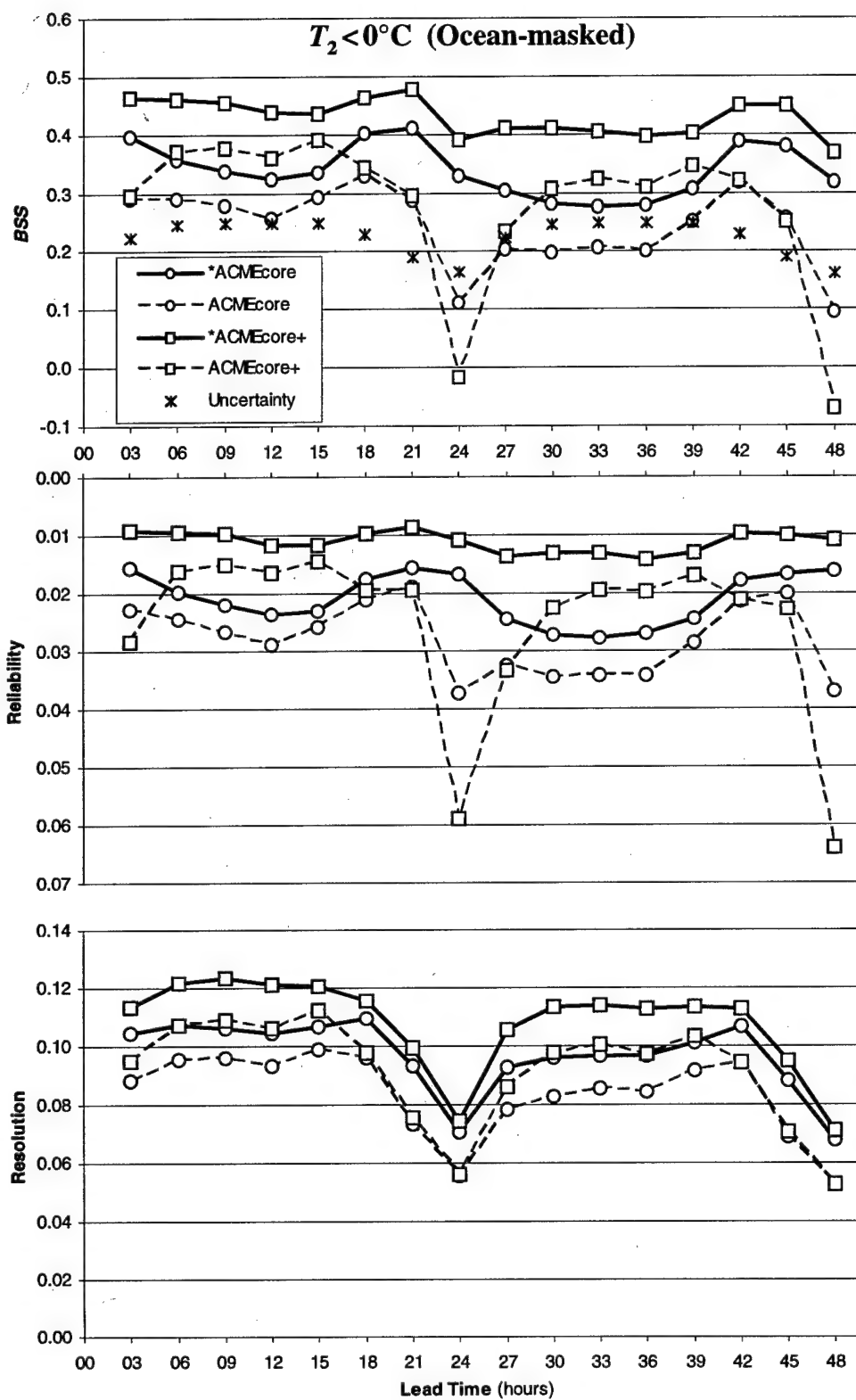


Figure 52. BSS and its components for FP of  $T_2 < 0^\circ\text{C}$  using ocean-masked data.

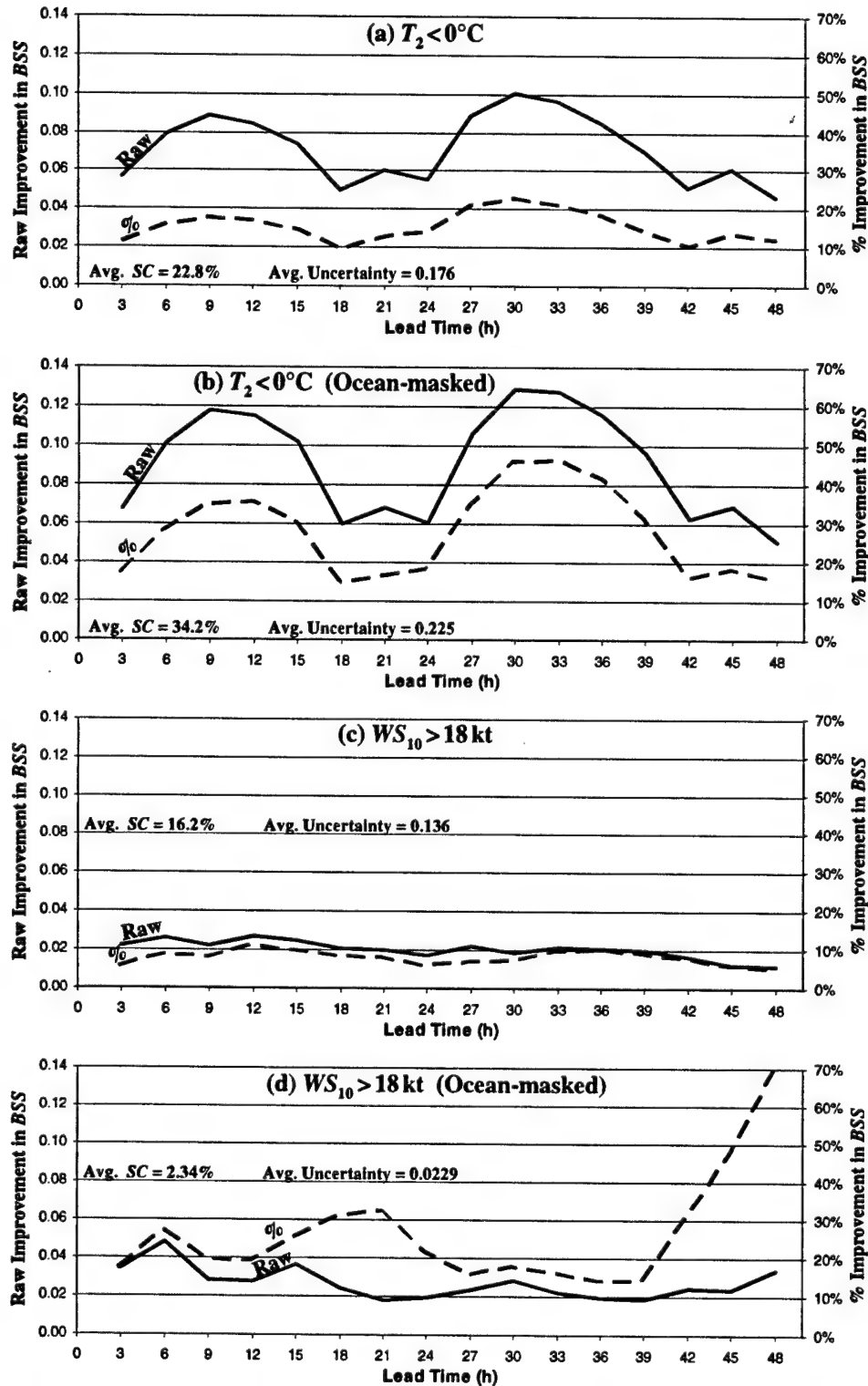


Figure 53. BSS improvement by \*ACME<sup>core+</sup> over \*ACME<sup>core</sup> for FP of the events: (a)  $T_2 < 0^\circ\text{C}$ , (b) Ocean-masked  $T_2 < 0^\circ\text{C}$ , (c)  $WS_{10} > 18 \text{ kt}$ , and (d) ocean-masked  $WS_{10} > 18 \text{ kt}$ . The average (over all lead times) SC and uncertainty for the events are indicated in each plot.

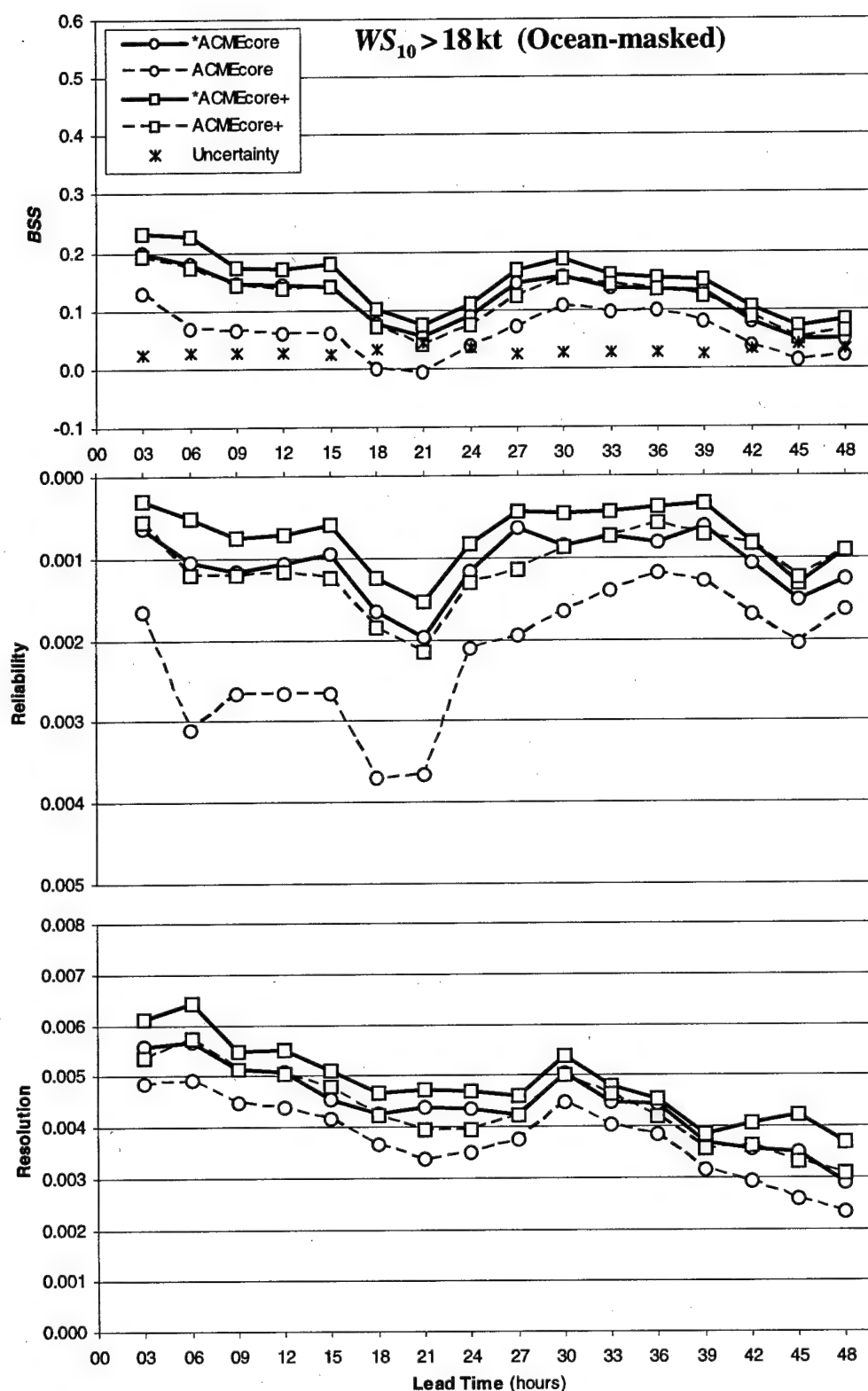


Figure 54. BSS and its components for FP of  $WS_{10} > 18$  kt using ocean-masked data.

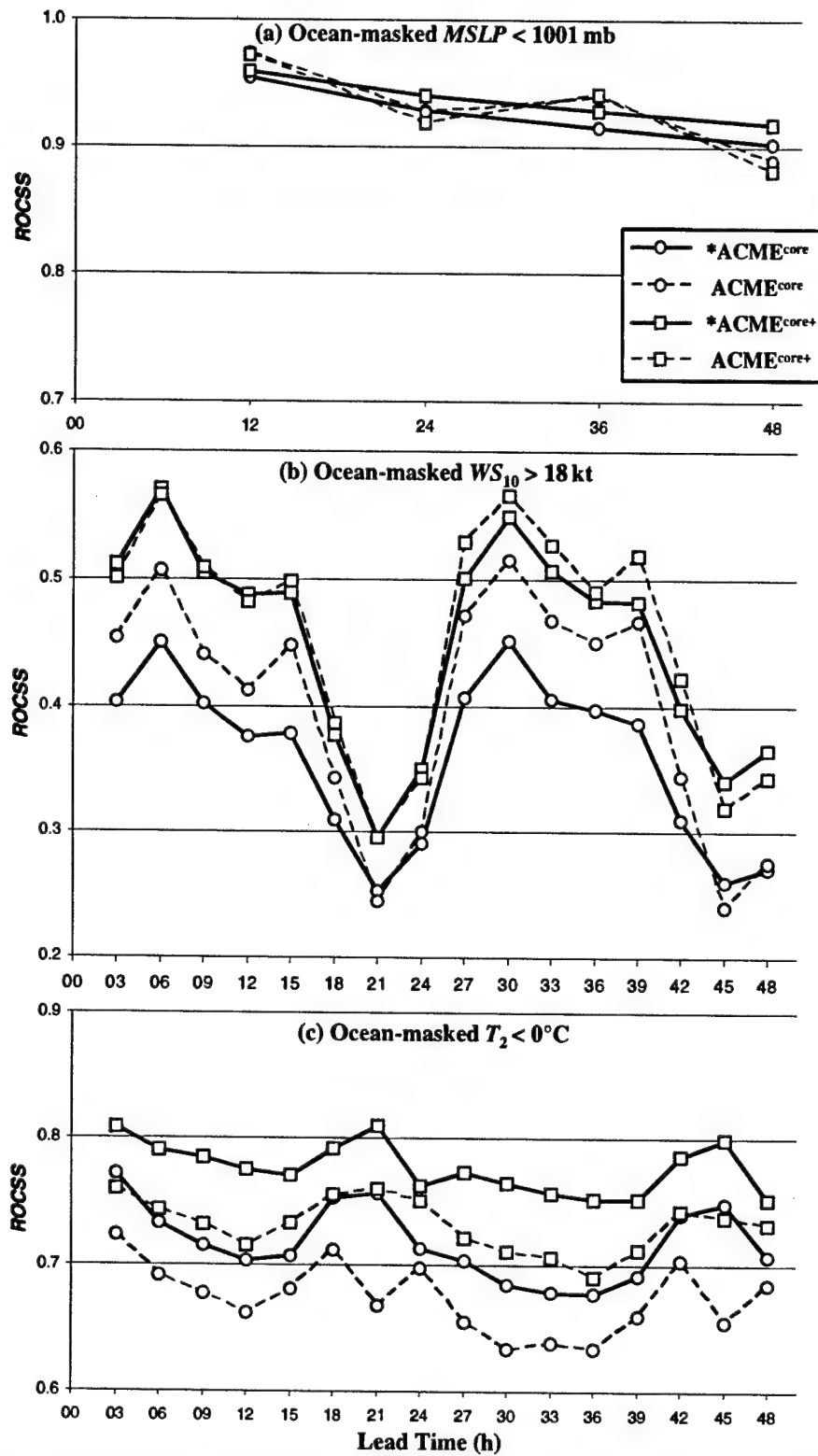


Figure 55. ROCSS for FP of (a)  $MSLP < 1001$  mb, (b)  $WS_{10} > 18$  kt, and (c)  $T_2 < 0^\circ\text{C}$ , using ocean-masked data.

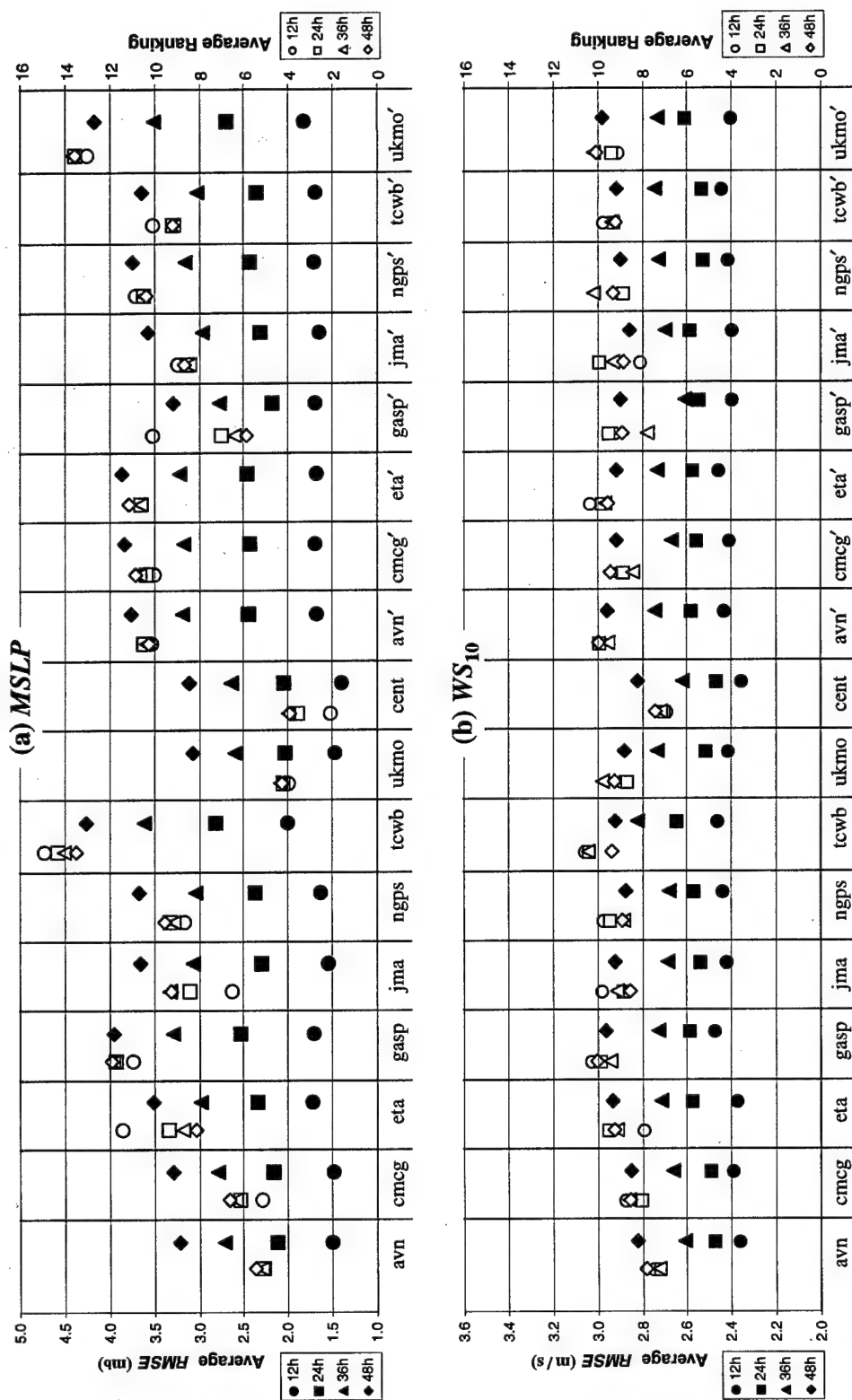


Figure 56. Deterministic skill comparison of ACME members by average *RMSE* and average ranking of (a) *MSLP* and (b) *WS*<sub>10</sub>.

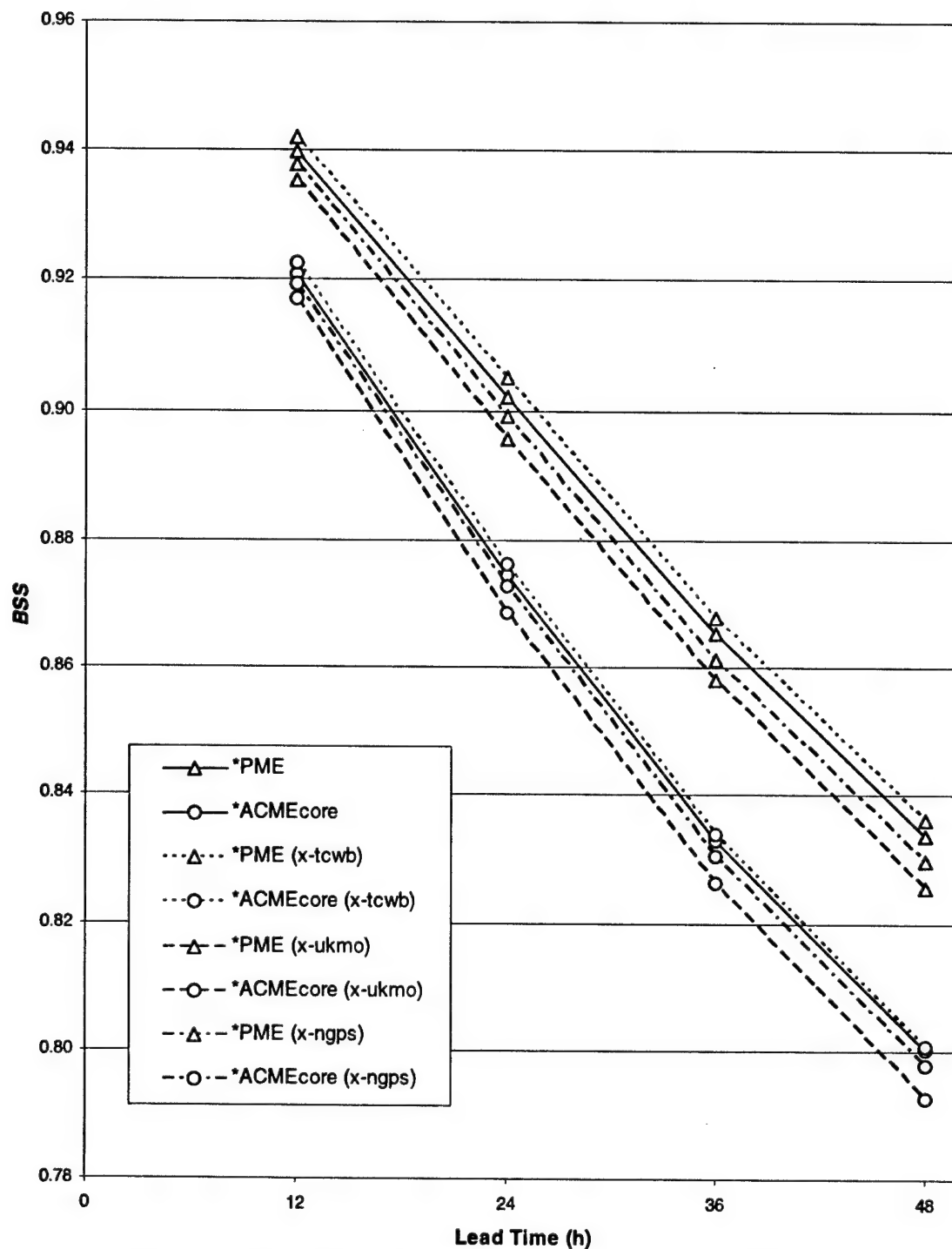


Figure 57. BSS for  $P(\text{MSLP} < 1001 \text{ mb})$  for regular \*PME (solid curve with triangles) and regular \*ACME<sup>core</sup> (solid curve with circles) compared to 7-member versions of the ensembles with tcwb withheld (dotted curves), ukmo withheld (dashed curve), and ngps withheld (dash-dot curve).



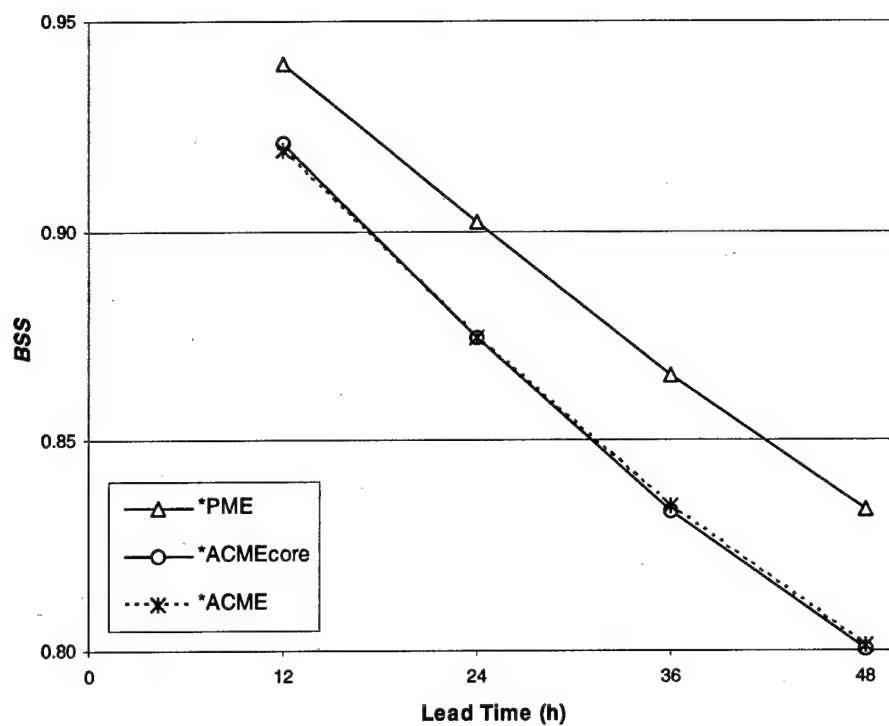


Figure 58. BSS for  $P(\text{MSLP} < 1001 \text{ mb})$  showing similarity between \*ACME<sup>core</sup> and \*ACME. \*PME results are included for reference.

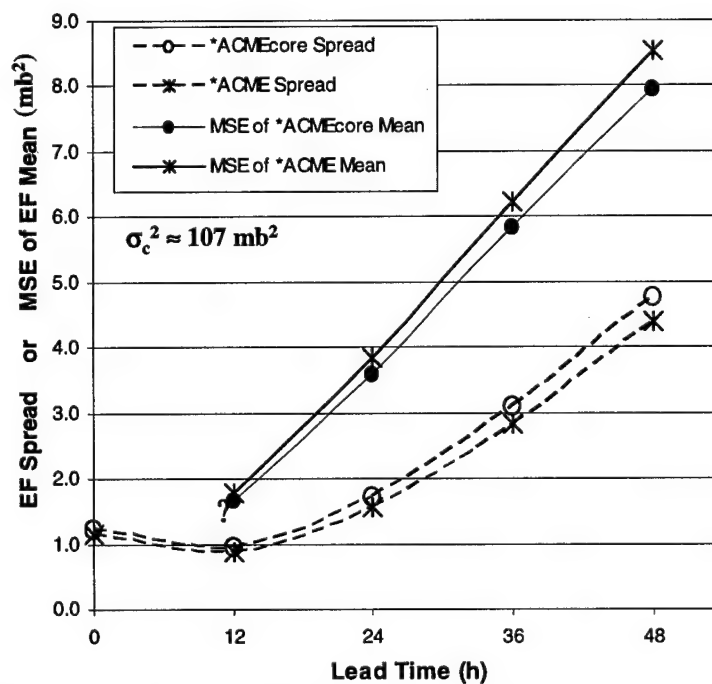


Figure 59. MSLP dispersion diagram for \*ACME<sup>core</sup> and \*ACME.

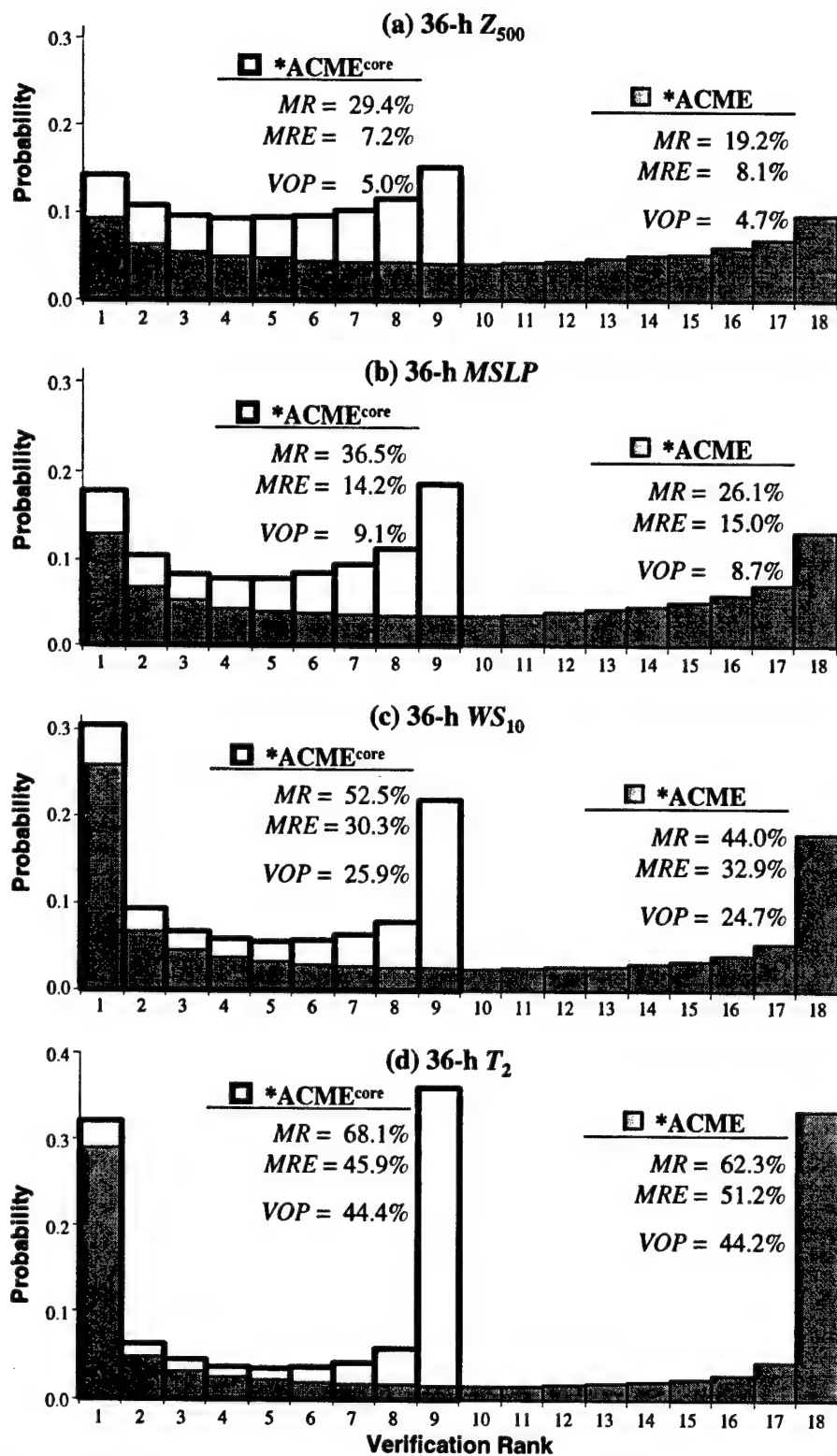


Figure 60. VRH comparisons between \*ACME and \*ACME<sub>core</sub>. Ideal MRs are 22.2% and 11.1% for \*ACME<sub>core</sub> ( $n = 8$ ) and \*ACME ( $n = 17$ ) respectively.

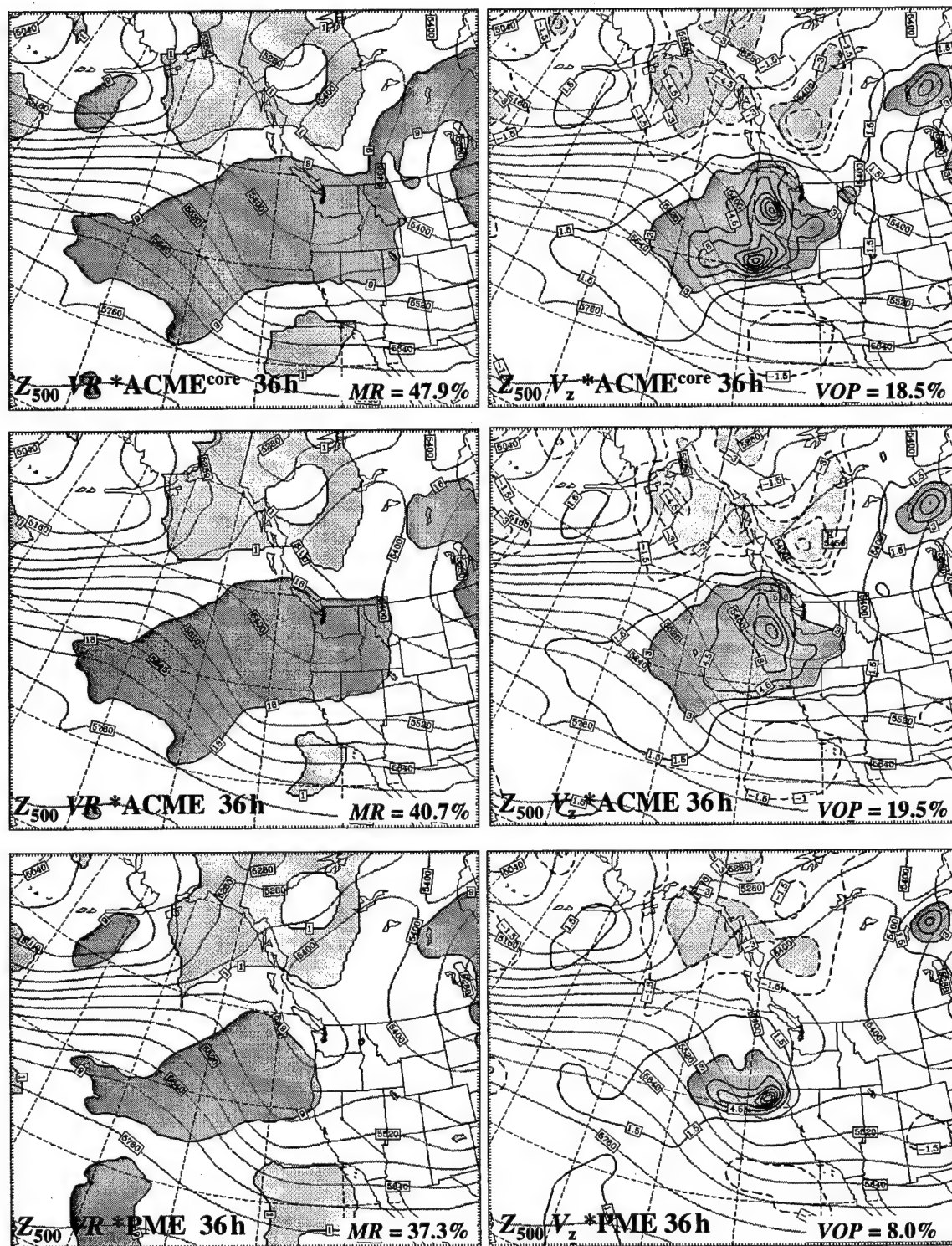


Figure 61.  $Z_{500}$  verification rank (VR) and  $V_z$  for 36-h lead time of the forecast initialized at 00Z, 20 Dec 2002 (the same case as in Figure 50).

Table 7. BSS data for  $MSLP < 1001$  mb.  $FP$  from all three SREF systems came from the same forecast cases so all systems share the same number of forecasts (# of Fcsts), number occurrences (# of Occ.), and  $SC$  of 0.2273 (or  $unc = 0.1756$ ). Bold, shaded values of  $res$ ,  $rel$ , and  $BSS$  indicate an improvement over the results of the SREF system to the left in the table.

FP	ACMEcore+					*ACMEcore+					*PME				
	# of Fcsts	# of Occ.	ORF	res	rel	# of Fcsts	# of Occ.	ORF	res	rel	# of Fcsts	# of Occ.	ORF	res	rel
0.00	832248	3224	0.004	41548.0	12.5	821579	2901	0.004	41141.4	10.2	815668	1070	0.001	41659.5	1.4
0.10	46885	6977	0.149	288.9	111.7	40642	3979	0.098	680.6	0.2	45431	2775	0.061	1255.3	68.8
0.20	26602	8621	0.324	249.1	409.5	23817	4481	0.188	36.5	3.3	25836	4086	0.158	123.6	45.2
0.30	16982	7910	0.466	965.8	466.8	16278	4996	0.307	103.2	0.8	16624	4640	0.279	44.6	7.3
0.40	14488	8319	0.574	1743.4	439.6	13877	5342	0.385	344.9	3.1	13461	5204	0.387	341.6	2.4
0.50	12983	8283	0.638	2189.7	247.2	12536	5948	0.474	765.8	8.2	12490	6157	0.493	881.4	0.6
0.60	13144	9351	0.711	3080.6	163.2	11961	6732	0.563	1346.5	16.5	12878	7638	0.593	1723.2	0.6
0.70	14995	11525	0.769	4393.3	70.5	14022	8863	0.632	2297.4	64.7	13934	9612	0.690	2980.8	1.4
0.80	22007	19073	0.867	8996.4	97.8	18654	14590	0.782	5742.4	6.0	18925	15911	0.841	7121.5	31.4
0.90	32655	30192	0.925	15876.3	19.7	32662	29341	0.898	14706.4	0.1	32833	30898	0.941	16726.8	55.4
1.00	158691	157403	0.992	92767.0	10.5	185652	183705	0.990	107855.7	20.4	183600	182887	0.996	108519.9	2.8
Total	1191680	270878		0.1444	0.0017	1191680	270878		0.1469	0.0001	1191680	270878		0.1522	0.0002
BSS				0.8124					0.8356					0.8655	

Table 8. Skill score comparison between \*ACME and \*ACME<sup>core</sup> at the 36-h lead time.

Event	BSS		ROCSS	
	*ACME	*ACME <sup>core</sup> Difference	*ACME	*ACME <sup>core</sup> Difference
$MSLP < 1001$ mb	0.8344	0.8330 0.0014	0.9611	0.9578 0.0033
$WS_{10} > 18$ kt	0.1939	0.2070 -0.0131	0.7580	0.7644 -0.0064
$T_2 < 0^\circ\text{C}$	0.4625	0.4669 -0.0044	0.8978	0.8999 -0.0021

#### IV. Summary

The basic premise of ensemble forecasting (EF) is that due to the inability to perfectly observe and model the atmosphere, the only complete way to predict its future state is to include the inherent uncertainty as part of the forecast process. In general, forecast uncertainty primarily results from errors in the analysis (i.e., model initial condition, IC) that grow nonlinearly (since the atmosphere is a chaotic system) during forecast integration. Depending upon the phenomenon and scale of interest, model error can also be a large source of forecast uncertainty.

EF is a method to incorporate both analysis and model uncertainty in the forecast process by using multiple runs of a numerical weather prediction (NWP) model where the IC and model of each ensemble member (i.e., individual model run) is varied according to their suspected uncertainty. The resulting set of solutions at any forecast lead time defines a probability density function (PDF) of future states of the atmosphere based on the uncertainties in the analysis and in the model. Given a large number of ensemble members, the forecast PDF is then a complete description of the future that widens with forecast lead time, reflecting the increase in forecast uncertainty. The challenge of EF is that, since analysis and model errors are not well understood, they are difficult to accurately represent in an ensemble system, making the ensemble's forecast PDF only an approximation. With a good approximation to the forecast PDF, there are many potentially beneficial EF products. In this research, we analyzed the skill of ensemble-based forecast probability (*FP*) for different events of interest (e.g., temperature less than freezing, or 10-m wind speed greater than 18 kt).

While there has been much success in approximating the forecast PDF for medium-range (2 – 10 days) ensemble forecasting (MREF), development of effective short-range (0 – 48 h) ensemble forecasting (SREF) has lagged behind for several possible reasons. First, the scale and parameters of interest in the short-range are less predictable so their errors may saturate too

quickly for an ensemble to be of use (i.e., a prediction based on climatology would have more value). Secondly, model uncertainty may have a larger impact on SREF parameters and it is difficult to represent such uncertainty in an ensemble since it is so poorly understood. Lastly, error growth is primarily linear in the short-range, which presents challenges for defining the ICs for SREF. For MREF, nonlinear error growth generates large, useful differences among ensemble members from almost any reasonable set of ICs.

The goal of this research was to explore the major issues of SREF and determine the effectiveness of real-time, mesoscale SREF using current capabilities and methods. A unique SREF test bed was built at the University of Washington by running the Fifth-Generation Pennsylvania State University–National Center of Atmospheric Research Mesoscale Model (MM5) using analyses from different operational forecast centers as ensemble ICs. The test bed included the following four distinct systems that produced 0 – 48-h forecasts (in real time, initialized daily at 00Z) over a large dataset of 129 cool season (Nov 2002 – Mar 2003) forecast cases over the Pacific Northwest (see Table 2, p 118).

- 1) **PME** (Poor Man's Ensemble, see Table 3, p 119): 8-member ensemble of low-resolution global models—a multimodel, multianalysis system.
- 2) **ACME<sup>core</sup>** (Core of the Analysis-Centroid Mirroring Ensemble): 8-member mesoscale ensemble running the MM5 (with 36/12-km nested domains and 32 levels) from the PME's analyses as initial/boundary conditions—a single-model, multianalysis system.
- 3) **ACME<sup>core+</sup>** (see Table 4, p 120): 8-member ensemble like ACME<sup>core</sup> but with variations to MM5—a perturbed-model, multianalysis system.
- 4) **ACME**: 17-member ensemble as an expanded version of ACME<sup>core</sup>—a single-model system with multi, centroid, and mirrored analyses.

For verification of these systems, the centroid analysis (mean of all 8 core analyses) was used on the outer, 36-km domain and the 20-km Rapid Update Cycle model analysis (RUC20) was used for the inner 12-km domain. A variety of statistical tools were used to analyze deterministic skill,

ensemble dispersion, and *FP* skill/value of 500-mb geopotential height ( $Z_{500}$ ), mean sea level pressure (*MSLP*), 10-m wind speed ( $WS_{10}$ ), and 2-m temperature ( $T_2$ ).

The first question we sought to answer was whether a group of independent analyses (i.e., multianalysis approach) provides a useful estimate of analysis error for SREF. Since the majority of forecast uncertainty for a synoptic-scale parameter comes from analysis error, skillful *FP* is only possible when ensemble ICs are able to represent that error well. The PME, ACME<sup>core</sup>, and ACME<sup>core+</sup>, which used the multianalysis approach, had positive *FP* skill for parameters with strong synoptic-scale influence (c.f., Figure 43, p 154 and Figure 54, p 167), implying that averaged over many forecast cases, analysis error was well represented by the spread of the analyses. However, the multianalysis approach is often hampered by high correlation among the analyses, which is discussed below.

For several reasons, the key to the success of our multianalysis approach for SREF is that the differences among the analyses are predominantly synoptic-scale. Ensemble ICs should include synoptic-scale differences since synoptic-scale errors are the largest errors generated by an analysis cycle. Secondly, it is the synoptic-scale errors that grow the largest so they must be included in ensemble ICs to consistently represent forecast uncertainty. Lastly, in the midlatitude cool season and over complex terrain, much of the mesoscale uncertainty is driven by the synoptic-scale error growth so small-scale errors may not need to be represented in the ICs.

The second question concerning analysis error was the possibility of expanding upon the core analyses to increase ensemble size beyond the limits imposed by the multianalysis approach (i.e., number of ensemble members equals the number of available analyses, which was 8 in our case). The purpose of the ACME system was to address this question and ameliorate the problems associated with small ensemble size. A small ensemble often does a poor job at representing the PDF from which the members are drawn, resulting in degraded *FP*.

ACME generated additional ICs from the information in the eight core analyses by using the difference between each core analysis and the centroid analysis as an estimate of the analysis error vector. The actual analysis error vector, which points in model phase space from the analysis to the true state, can never be known. The components of each of our eight estimates of the analysis error vector were made up of the grid point by grid point difference between the centroid analysis and each core analysis of all state variables at all model levels. Each estimated analysis error vector contained structural information (e.g., variation in position of a long wave trough) on the likely analysis errors besides error magnitude. To generate another possible IC, an analysis error vector was added back onto the centroid analysis, producing a mirror of an original core analysis about the centroid analysis. Lateral boundary conditions (LBCs) were handled in exactly the same manner.

The mirroring process of ACME produced an additional 9 ensemble members since the centroid analysis and each of the 8 mirrored ICs/LBCs was used to run MM5. We found that these additional members yielded unique solutions that were generally on a par with the forecasts from the core analyses (c.f., Figure 56, p 169). We also found that there was considerable variability in skill among the members, including among the core members, which is generally considered to be a detrimental attribute for an EF system.

In strict EF theory, it is required that all members be independent and equally likely so that they can be considered random draws from the forecast PDF. However, this work demonstrated that an effective ensemble can be made up of unequally skilled members. A member with lower average skill can add value to an ensemble as long as it occasionally performs well. Only a member that performs poorly the majority of the time can degrade the skill of an ensemble (c.f., Figure 57, p 170).



Comparing ACME and ACME<sup>core</sup>, we found that ACME unfortunately did not demonstrate an improvement in skill commensurate with the increase in ensemble size from 8 to 17. However, ACME was better able to encompass (i.e., completely surround) the verification compared to ACME<sup>core</sup>, so the additional ACME members did provide valid forecasts not produced in ACME<sup>core</sup>. To explore how ACME could appear to be of no additional value for *FP* but at the same time be valuable for encompassing truth more often, we developed a new analysis tool called the standardized verification ( $V_Z$ ).

For highly reliable *FP*, the verification should appear to be drawn from the PDF represented by the ensemble members—an objective termed ‘statistical consistency’.  $V_Z$  tests for failure to meet statistical consistency by subtracting the ensemble mean from the verification at each grid point and then dividing by the ensemble standard deviation ( $s$ ), thus translating the verification into  $s$  units. A large  $V_Z$  value (chosen as  $> 3$  or  $< -3$ ) indicates that the verification was an outlier with respect to the ensemble PDF (i.e., truth “got away” from the ensemble and was not a random draw from the ensemble’s PDF). Comparing plots of  $V_Z$ , we found that ACME and ACME<sup>core</sup> shared the same areas where the verification was an outlier (c.f., Figure 61, p 173). Furthermore, the areas where ACME<sup>core</sup> had failed to encompass the verification, but ACME did, were actually areas where  $|V_Z| < 3$  for ACME<sup>core</sup> (i.e., verification was not an outlier). Therefore, the apparent gain made by ACME in encompassing truth was of little value since it did not correct the problems (i.e., where truth got away) of ACME<sup>core</sup>, which is why ACME provided no improvement in *FP* skill.

Our conclusion concerning use of independent analyses as ensemble ICs is that, while the differences among the analyses do represent analysis error well on average, there are often occasions when the analyses share similar errors and are too highly correlated. For example, over a poorly observed area of the Pacific Ocean all the analyses may omit a developing short wave,

which leads to a large region in the forecast fields where truth gets away from the ensemble since none of the members would contain the ensuing cyclone. The mirroring method was successful at producing more valid samples from the forecast PDF of ACME<sup>core</sup>, but ACME did not improve *FP* skill because it only provided more samples from a deficient PDF. The additional members of ACME could only vary the position of short waves represented in ACME<sup>core</sup> but could not produce short waves that were entirely missed by all the analyses. A valuable outcome from the ACME system was that the MM5 run from the centroid analysis displayed the best overall deterministic performance among the individual ensemble members. This likely means that the centroid analysis is the best representation of synoptic-scale truth, although it does tend to smooth out structures.

Another major question we researched was by how much and by what means do model deficiencies (both stochastic and systematic error) impact SREF skill and value? Our results showed conclusively that model deficiencies do play a significant role in SREF. Stochastic errors (i.e., random model errors) are a large source of uncertainty and must be accounted for within a SREF system in order to maximize utility, particularly for mesoscale, sensible weather phenomena. Systematic errors (i.e., model biases) are clearly not part of the forecast uncertainty but are a large part of the forecast error and can seriously degrade ensemble performance if not corrected.

To eliminate the bulk of the systematic error, we applied a simple grid-based, 2-week, running-mean bias correction to each member separately. This approach was based on findings that biases are predominantly linear and dependent on location, forecast lead time, weather regime, and ensemble member (c.f., Figure 24 – Figure 26, p 102 – p 104). To demonstrate a method for real-time application, we used the previous two weeks as training data for the bias correction of each forecast cycle. A 2-week training period was used to: 1) capture the short

timescale variability in bias that arises from shifts in weather regime, and 2) obtain a reasonably sized sample of data at each point. We found that forecast bias can be influenced by both the model and the analysis, so fixing the IC and model for the members in our ensemble systems made bias correction more effective.

The larger and more consistent the bias in a parameter, the more improvement was realized from the bias correction (c.f., Figure 28 – Figure 34, p 106 – p 112). The PME members generally had lower bias compared to the ACME system members since the large-scale models have lower resolution and are better tuned. A mesoscale model produces more bias as it attempts to represent smaller-scale phenomena with additional parameterizations.

Bias correction benefited SREF by greatly improving *FP* skill by: 1) improving reliability by adjusting the mean of the ensemble's PDF to match the mean of the verification's PDF, and 2) improving resolution by narrowing the ensemble's PDF where members had opposing biases. Figure 43 (p 154) shows how bias correction improved the performance of ACME<sup>core</sup> by 6 h, which is significant for a short-range forecast. An additional benefit of bias correction was that analyzing bias-corrected results led to firmer conclusions, such as the importance of accounting for stochastic model error. For example, in Figure 52 (p 165), it is only after bias correction that ACME<sup>core+</sup> stands out as superior to ACME<sup>core</sup> at all lead times.

The ACME<sup>core+</sup> system was designed to account for model uncertainty and explore how inclusion of model diversity affects a SREF on the mesoscale and for sensible weather. ACME<sup>core+</sup> (see Table 2, p 118) applied the perturbed-model strategy in which the members used the same ICs as ACME<sup>core</sup> but each was given a unique version of MM5. The goal of this approach was to generate large and realistic dispersion that represented the model uncertainty of MM5. Model perturbations consisted of different combinations of physics options (planetary boundary layer, cloud microphysics, cumulus, and radiation schemes) and randomly perturbed

surface boundary parameters (SBPs) (sea surface temperature, moisture availability, albedo, and roughness length).

Comparing  $ACME^{core}$  and  $ACME^{core+}$ , we found that inclusion of model diversity dramatically increased ensemble spread, which improved statistical consistency but still fell well short (c.f., Figure 47c & d, p 158). (For statistical consistency, ensemble spread must match the mean square error of the ensemble mean when averaged over many forecast cases.) So while  $ACME^{core+}$  was able to improve *FP* skill (discussed below), there is still room for improvement in our mesoscale SREF methodology. This analysis also revealed a dramatic lack of error growth for  $T_2$  (Figure 47d) but the error was not saturated in the short range. Since ensemble spread was such a small fraction of forecast error for  $ACME^{core}$  (in which all members shared the same version of MM5), we concluded that for our dataset the error in  $T_2$  is dominated by model error. In other words, analysis error and the error growth it produces contributes very little to the observed  $T_2$  forecast errors—a result completely different from that for a variable with strong synoptic-scale influence, such as  $WS_{10}$ . That finding led to a general conclusion: the relative influence of analysis and model uncertainty for SREF is greatly dependent upon the scale and variable of interest.

$ACME^{core+}$  did display greatly improved (in both reliability and resolution) *FP* skill over  $ACME^{core}$ , revealing that model errors are a large part of the forecast error at the mesoscale and can be at least partly represented by the perturbed-model approach (c.f., Figure 52, p 165 and Figure 54, p 167). Unlike bias removal that improves skill by narrowing the forecast PDF away from values where the verification is unlikely to occur, including realistic model diversity improves skill by widening the forecast PDF toward values where the verification may occur. We also confirmed that including model diversity is more important near the surface over land where model parameterizations have the greatest influence (c.f., Figure 53, p 166).

Further study on the issue of representation of model uncertainty was performed by comparing the multimodel approach of the PME to the perturbed-model approach of ACME<sup>core+</sup>. It was expected that the PME would exhibit greater dispersion since the model differences among the PME members are likely much greater. We found that the PME was actually slightly overdispersive (c.f., Figure 47a & b, p 158) and performed much better on the synoptic scale compared to ACME<sup>core+</sup> (c.f., Figure 51, p164). Just as the differences between model options in ACME<sup>core+</sup> appear to represent model error to some degree, the large differences among the PME members' models can skillfully represent model error. Furthermore, the greater model diversity within PME makes it more skillful than ACME<sup>core+</sup>. The downside of the PME is that it does not include the desired information on the mesoscale—the reason for the implementation of ACME<sup>core+</sup>.

We proposed a two-part strategy for improving the low dispersion problem of ACME<sup>core+</sup>, which should also improve *FP* skill. First, the MM5 forecast of each ACME<sup>core+</sup> member should be periodically nudged toward the large-scale model from which it was forced, thus improving the large-scale dispersion. Besides greater model diversity, the PME produces more dispersion than ACME<sup>core+</sup> because the PME grows the large-scale errors globally whereas a mesoscale ensemble reduces error growth by running on a limited-area domain, even with updated lateral boundaries. The second part of the solution deals with small-scale error growth. We found that our 12-km domain was able to produce greater ensemble spread compared to the 36-km domain since finer model resolution is able to capture variability on smaller scales of motion (c.f., Figure 49b). We proposed that further increasing the grid resolution should produce higher, more accurate dispersion among the ACME<sup>core+</sup> members and thus more highly skilled *FP*. Higher resolution would also have the added benefit of reduced reliance on physical parameterizations so their errors would no longer have to be approximated.

A final comment on the issue of representing stochastic model error in SREF concerns how much improvement we should expect from more thorough error representation. While it is clear that there is value in either the multimodel or perturbed-model approaches, their chief limitation is the use of gross model differences (either by differences between models or model options) to approximate model error rather than rigorously perturbing all parameterizations individually. How much more value could be realized from SREF by perturbing the model more rigorously? We speculate that such an effort would not be worth the cost since we may never truly understand many aspects of model error and therefore never be able to perturb them rigorously. It may be more beneficial to SREF to focus on improving the mesoscale model to reduce the uncertainty within the model.

In analyzing our MM5 SREF results, it became clear that there are significant deficiencies in current mesoscale modeling. For example, the error in  $T_2$  is about the same in the first few hours of the forecast as it is at the 48-h lead time, which reveals the models inability to represent surface and boundary layer effects. Analysis of SREF may help to identify the most deficient aspects of the model. Dispersion diagrams such as Figure 47 (p 158) reveal poor model dispersion and point out where model options may be unable to represent certain atmospheric behaviors. Plots of standardized verification such as Figure 61 (p 173) identify structures that are not represented by the forecasts and may be traced to model or analysis deficiencies.

In closing, this research was not an attempt to build an ideal SREF but rather an opportunity to realize most of the potential SREF benefits by employing sound methods that are currently computationally feasible. Detailed analysis revealed that while there are limitations to SREF, there is value in mesoscale SREF even with today's capabilities. Intercomparison of our different systems yielded answers to basic SREF issues that apply to the development of more optimal SREF systems in the future.

## Appendix I: EF Statistical Toolbox

This appendix presents statistical techniques for evaluating the quality of an EF system, which is not a straightforward matter. Similar challenges as faced by deterministic forecast verification are present for evaluation of EF, such as verification on the appropriate scales, differences between observation-based and model analysis-based verification, interpolation of data from observation locations to model grid points or vice versa, and errors or biases in the verification itself. For EF, there is the additional problem of verifying a stochastic-type forecast with deterministic observations since stochastic observations are generally not available. The tools explained here were designed to meet these challenges, but each has unique strengths and weaknesses. When used collectively these tools represent a fair and thorough means to evaluate and compare EF systems.

There are two general types of EF statistical evaluation tools: *consistency* tools and *utility* tools. A consistency tool evaluates whether the verification can be considered a random sample from the PDF defined by the ensemble members. This is a necessary condition for the ensemble to properly represent the forecast uncertainty and is often termed *statistical consistency*. A utility tool evaluates whether or not an ensemble can produce useful information for a particular user or users in general. An EF system can be statistically consistent but yet of little value. For example, highly reliable forecast probability (*FP*) could be provided simply by the climatologic norm but such forecasts would not be able to distinguish between events and nonevents (i.e., possesses no resolution).

### A. Dispersion diagram

A dispersion diagram is a consistency tool that displays how well the mean square error (*MSE*) of the ensemble mean matches the ensemble variance. Besides statistical consistency, it

also reveals the predictability error growth of the variable being examined. This diagram and the Error Variance Diagram are thoroughly detailed in section I.B.1 (page 13) so will not be covered here.

## B. Verification Rank Histogram

The verification rank histogram (VRH) is a consistency tool based on Anderson's (1996) binned probability ensemble technique. It is a useful tool for visualizing statistical consistency and the dispersive character of an ensemble. Construction of a VRH for a parameter such as 850-mb temperature ( $T_{850}$ ) in Figure 62 begins by pooling a verification value at one location with the forecast values from an  $n$ -member ensemble, followed by sorting of the  $n+1$  values from least to greatest. The resulting rank (i.e., ordered position among the  $n+1$  values) of the verification is recorded over many such trials (over space and/or time) to build a histogram of the number of occurrences within each rank. Dividing the total verifications that occurred in each rank by the total number of trials gives the probability that the verification occurred within each rank.

An example of a hypothetical trial, just one datum for the construction of a Figure 62, is detailed in the "realistic forecast" in Table 9 and Figure 63. With eight forecasts, there are nine possible ranks for the verification, so if the observed  $T_{850}$  is  $2.1^{\circ}\text{C}$  then a verification rank of 6 is recorded for this trial. In the event that the verification exactly equals one or more of the EF forecasts, the rank is randomly assigned among its possible values (Hamill and Colucci, 1997). (E.g., if the observed  $T_{850}$  in the example was  $-0.18^{\circ}\text{C}$  then verification rank is randomly assigned to rank 3 or rank 4.)

For a very large number of trials, a well calibrated EF produces a uniform VRH if the verification is a random draw from the same PDF as the EF's forecast PDF. In other words, on average the verification should have an equal chance of occurrence in each rank equal to  $1 / (n + 1)$  (Anderson, 1996). This may seem counterintuitive at first given that the widths of the



ranges of possible verification values within each rank can vary quite a lot. Consider the idealized forecast in Table 9 and corresponding plots in Figure 63a & b. From a quantile point of view, the eight forecasts are uniformly spread, so, while each of the nine possible verification ranks has different sized ranges of the random variable  $T_{850}$ , the probability of occurrence in each rank is still  $1/9$ .

However, in the realistic example (Figure 63c and d) it is clear that, besides different sized ranges of  $T_{850}$ , there is also unequal probability among the nine possible verification ranks. For this single trial, the verification will most likely occur in rank 3. But over many trials, the average probability of occurrence in each rank will equal  $1/9$ . This is why the principle of VRH uniformity for a well-calibrated EF applies only to very large amounts of data.

When the verification's PDF is quite different from the EF's forecast PDF, the probability of occurrence among ranks will not be uniform on average. Figure 62 shows a u-shaped VRH, commonly found in EF, which indicates the verification has a greater variance compared to the forecast PDF since the verification occurs too often in the extreme ranks. A u-shaped VRH may also indicate weak dispersion of the ensemble members since truth is too often not encompassed.

Note that a uniform VRH is a necessary but not sufficient condition for an EF system to be considered well calibrated. It is possible for problems of an EF system to be camouflaged by various aspects of the forecast and verification data (Hamill, 2001). For example, Hamill (2001) demonstrated how an overdispersive EF system that also has a conditional bias (positive at times and negative at other times) can produce a uniform VRH. Furthermore, a uniform VRH is not a measure of an ensemble's skill since uniformity could simply be achieved by forcing spread toward climatology, which would reduce skill.

A factor often analyzed from a VRH is the *missing rate (MR)*, which is the total percentage of verifications that occurred in the outer ranks (i.e., rank 1 and rank  $n+1$ ):

$$MR = 100 \left( \frac{N_1 + N_{n+1}}{M} \right) \quad (33)$$

where  $N_x$  is the number of verifications that occurred in a rank  $x$ , and  $M$  is the total number of trials. A  $MR$  greater (less) than the statistically consistent value of  $2/(n+1)$  provides a quantitative evaluation of the underdispersion (overdispersion) of an EF system and also the ability of the EF to encompass truth. In comparing the  $MR$  between ensembles of difference size, it is better to compare the missing rate error ( $MRE$ ) since the statistically consistent value of the  $MR$  depends on  $n$ :

$$MRE = 100 \left( \frac{N_1 + N_{n+1}}{M} - \frac{2}{n+1} \right) \quad (34)$$

### C. Standardized Verification

As described above, the  $MR$  or  $MRE$  only reveals an ensemble ability to encompass truth and can not answer the more important question of an ensemble's ability to portray truth. (Recall our definition that the verification is portrayed if it occurs within three standard deviations from the mean of the EF's approximate forecast PDF.) A high value of  $MR$  could be associated with few or many verification values not portrayed depending upon the shape of the PDF tails involved. Also, a low  $MR$  (generally indicating overdispersion) does not necessarily mean that the verification is being portrayed too often.

A way to measure an ensemble's ability to portray truth follows the statistical calculation of the standard normal random variable which transforms the value of a variable into units of standard deviation (Devore, 1995):

$$V_z = \frac{V - \bar{e}}{s} \quad (35)$$

where  $V_Z$  is termed the standardized verification,  $\bar{e}$  is the ensemble mean,  $V$  is the verification value, and  $s$  is the ensemble standard deviation, all at a single grid point. This provides an excellent tool to determine when and if the verification is an outlier and not portrayed by the ensemble. Furthermore, by plotting  $V_Z$  for a single forecast case, it is possible to explore how truth gets away from an ensemble by revealing any structure to the regions where the verification occurs beyond  $3s$ . Note also that  $V_Z$  carries a sign to indicate the direction (+, high; -, low) of the verification in relation to the ensemble mean. Note that while  $V_Z$  can reveal where an ensemble has failed its primary goal of portraying the truth,  $V_Z$  can not show anything concerning whether the ensemble is overspread.

As another check for statistical consistency, we could also calculate an average  $V_Z$  over many points:

$$\overline{V_Z} = \frac{1}{M} \sum_{m=1}^M \frac{|V_m - \bar{e}_m|}{s_m} \quad (36)$$

where  $M$  is the number of data points being verified, and the  $m$  subscripts reference a single grid point. However, while the expected value of  $\overline{V_Z}$  is around 1.0, it also depends upon ensemble size and the shape of the distribution. Therefore  $\overline{V_Z}$  can not easily be used to measure the statistical consistency of an ensemble.  $V_Z$  is not truly standardized to 1.0 for our purposes.

An overall measure that is useful will be termed the verification outlier percentage,  $VOP$ :

$$VOP = \frac{100}{M} \sum_{m=1}^M \left\{ \begin{array}{ll} 0: & 3s_m \geq |V_m - \bar{e}_m| \\ 1: & 3s_m < |V_m - \bar{e}_m| \end{array} \right\} \quad (37)$$

Basically this finds the average percentage of the data pairs in which the verification is not portrayed by the ensemble. That is, if the verification falls beyond  $3s$  from the mean on either side, we call it an outlier with respect to the EF. For a normal PDF, outliers beyond  $3s$  are rare but are still expected to occur ~0.3% of the time. Therefore, the amount that  $VOP$  exceeds 0.3%

is a measure of how much truth gets away from the ensemble. However, for the same reason that  $V_Z$  is not actually a standardized measure, that rule is only a rough guide.

It is more useful to compare the  $VOP$  between ensemble systems as a relative measure of each system's ability to portray truth. This is superior to comparing the missing rate because a similar value of missing rate can have a variable  $VOP$ . For example, we found the missing rate for  $Z_{500} *ACME^{core+}$  to be 27.50% and 27.52% for both the 36- and 48-h lead times respectively, but  $VOP = 4.16\%$  at 36 h and 3.85% at 48 h. The  $VOP$ s indicate that  $*ACME^{core+}$  was better able to portray the truth at 48 h while the  $MR$  could not because of its limitations.

#### D. Brier Score and Brier Skill Score

The Brier score ( $BS$ ), essentially a mean square error measure for  $FP$  (Wilks, 1995), measures the accuracy of a set of  $FP$ s for the same event. With a large number of such forecasts and corresponding verifying observations, the  $BS$  is calculated as (Equation 7.22, Wilks, 1995):

$$BS = \frac{1}{N} \sum_{i=1}^N (FP_i - OBS_i)^2 \quad (38)$$

where  $N$  is the total number of forecasts/observation samples, and  $FP_i$  is the forecast probability of the  $i^{th}$  sample.  $OBS_i$  equals 1 if the event occurred for the  $i^{th}$  sample, and 0 if it did not occur. Therefore,  $BS$  varies between 0 (perfectly accurate) and 1 (totally inaccurate).

The  $BS$  is very useful for comparing the relative skill of two sets of probability forecasts (e.g., forecasts from the  $ACME^{core}$  vs. forecasts from  $ACME^{core+}$ ). A more explicit measure for a single set of forecasts, called the Brier skill score ( $BSS$ ), can be made by comparing the  $BS$  to  $BS_{clim}$ , which comes from forecasts based on the climatologic probability of occurrence.

$$BSS = \frac{BS - BS_{clim}}{BS_{perfect} - BS_{clim}} = 1 - \frac{BS}{BS_{clim}} \quad (39)$$

since  $BS_{\text{perfect}} = 0$  (Equation 7.23, Wilks, 1995). Equation (39) gives the amount of improvement over the climatologically based forecasts. The  $BSS$  is a utility tool where a value of 1.0 indicates perfect forecasts and a value of 0.0 or less indicates a worthless forecast.

### E. Reliability Diagram

A reliability diagram is a graphic display of the  $BS$  created by binning the continuous  $FP$  values into  $I$  discrete, contiguous bins of probability, then plotting the  $FP$  at the center of each bin ( $FP'$ ) against the corresponding observed relative frequency ( $ORF'$ ). The  $BS'$  can then be calculated through decomposition into *reliability* (*rel*), *resolution* (*res*), and *uncertainty* (*unc*).

$$BS' = \underbrace{\frac{1}{M} \sum_{i=1}^I N_i (FP'_i - ORF'_i)^2}_{(rel)} - \underbrace{\frac{1}{M} \sum_{i=1}^I N_i (ORF'_i - SC)^2}_{(res)} - \underbrace{SC(1 - SC)}_{(unc)} \quad (40)$$

where  $M$  is the total number of forecasts/observation data pairs for the event,  $i$  is the index for the  $I$  bins,  $N_i$  is the number of forecasts within the  $i^{\text{th}}$  bin,  $FP'_i$  is the forecast probability at the center of each bin,  $ORF'_i$  is the observed relative frequency for the forecasts in bin  $i$ , and  $SC$  is the sample climatology (Equation 7.28, Wilks, 1995). Note that Equation (40) is an approximation to Equation (38) because of the binning of the forecast probabilities. If  $FP_i$  values were rounded to  $FP'_i$  values for use in Equation (38), or if  $I \rightarrow \infty$  in Equation (40), then the two equations would be equivalent.

The reliability diagram plots  $ORF'_i$  vs.  $FP'_i$  so a perfectly reliable forecast follows a line of slope = 1.0 starting at the origin. (E.g., for the set of all cases in which 20% chance of occurrence was forecast, the event should be observed to occur for 20% of those cases.) The *rel* term is a measure of the distance away from the perfect forecast line weighted by the number of forecasts at each  $FP'_i$ . Better forecasts result in a smaller *rel* and thus a  $BS'$  closer to zero. Note that a forecast based on the climatologic probability has perfect reliability ( $rel = 0.0$ ).

Figure 64 is a detailed example of a reliability diagram from Eckel (1998) built from the data in Table 10. The event forecast was 24-h cumulative precipitation  $> 0.25$  inch at the 36-h forecast lead time. Since  $M = 11$ , the original forecast probabilities are, in effect, rounded to the nearest 10%, making  $FP_i'$  bins of 0.0% – 4.9%, 5.0% – 14.9%, ..., 95.0% – 100%. The histogram in Figure 64 is a display of the relative frequency of usage of the forecast probabilities (i.e., how many forecasts were made within each  $FP_i'$  bin).

The  $SC$  is the overall frequency of occurrence of the event. The resolution term is a measure of distance away from the climatologic probability forecast (dashed line labeled zero resolution) weighted by the number of forecasts at each  $FP_i'$ . Better forecasts result in a larger  $res$  and thus a  $BS$  closer to zero. Resolution is a measure of the forecast's ability to discriminate between occurrence and nonoccurrence of the event. It is possible then to improve the reliability of a forecast system by increasing its spread toward the climate PDF (Evans, 2000). But this does not improve the system's quality since  $res$  would decrease. Note that a set of forecasts based on the climatologic probability of occurrence has the worst possible resolution ( $res = 0.0$ ) since for such forecasts,  $ORF = SC$ .

The uncertainty term is determined by the  $SC$  and thus independent of the forecasts (Figure 65). It can be thought of as a measure of how easy it is to forecast the event in question. The highest possible  $unc$  of 0.25 (most difficult to forecast) is associated with an event that occurs half of the time on average (i.e.,  $SC = 50\%$ ). An event that rarely occurs or frequently occurs has a lower  $unc$  (easier to forecast) with a minimum of 0.0 when  $SC = 0.0\%$  or  $100\%$ .

The  $BSS$  can be computed from a reliability diagram by applying the fact that for a climatologically based forecast the  $res$  and  $rel$  terms are both zero, as described above. Substituting the  $BS'$  from Equation (40) into Equation (39), we get (Equation 7.29, Wilks, 1995):

$$BSS' = 1 - \frac{rel - res - unc}{0 - 0 + unc} = 1 + \frac{res - rel}{unc} - \frac{unc}{unc}$$

$$BSS' = \frac{res - rel}{unc} \quad (41)$$

Therefore, for a point in a reliability diagram to contribute positive skill, it must have  $rel < res$ .

This requirement defines a *skill zone* as the shaded region in Figure 64. Forecasts that exhibit an overall negative  $BSS'$  performed worse than a simple climatologically based forecast.

Note that the main difference between  $BSS$  and  $BSS'$  is in the choice of what is used as the climatologically based forecast. The  $BSS$  can be calculated with respect to the  $BS_{clim}$  from a long-term climatologic forecast, while the  $BSS'$  uses the short term average chance of occurrence over just the dataset (i.e., the  $SC$ ). In this respect, the  $BSS'$  is a more stringent score since, for a limited dataset, the  $SC$  reflects the climatology of the sample and should therefore produce a better average  $FP$  compared to the long-term climatologic forecast. Furthermore, since the long-term climatology for an event is often difficult to obtain, normally only  $BSS'$  is computed. (Note: In the body of this dissertation we drop the prime notation since the only  $BSS$  employed is from Equation (41).)

Interpreting a reliability diagram can be tricky. A primary concern is the sample size since to be confident in the diagram, there should many samples (perhaps minimum of about 50) within each  $FP'_i$  bin. Next, the relative sampling among the  $FP'_i$  bins (the histogram in Figure 64) needs consideration since that shows how the  $res$  and  $rel$  components are weighted. Obviously, it is desirable for data points to fall in the skill zone, but some points with small weight may fall outside so that an overall positive  $BSS'$  results.

The two basic curves often observed in a reliability diagram of an EF, the  $S$  and the reverse  $S$ , are a result of the basic dispersion characteristics of the EF. Points above the perfect line are

associated with underforecasting while points below are overforecasts. So the reverse *S* means that the EF overforecasts the higher probabilities and underforecasts the lower probabilities. It can be seen that this corresponds to an underdispersive EF by considering the u-shaped VRH. When an underdispersive EF gives a *FP* of 10% for some event threshold, the actual chance of the verification falling above the threshold is much higher (i.e., large probability in the outer rank) so this is an underforecast. The reverse holds for the case of a high *FP*. The entire scenario flips for the *S*-shaped reliability diagram curve so that it corresponds to an overdispersive EF. While the point of inflection for these curves can be pushed toward one end of the diagram, it will still be identifiable as either the *S* or the reverse *S*.

#### F. Relative Operating Characteristic

The relative operating characteristic (ROC) is a verification tool that employs signal detection theory, a technique designed to evaluate binary-type forecasts in which forecasts are restricted to a "no" (i.e., *FP* = 0%) or a "yes" (i.e., *FP* = 100%). The initial step in computing the ROC is to reduce the full probabilistic information from an EF down to binary-type forecasts for the event.

For example, say an EF forecasts a 37% chance for the event that 24-h cumulative precipitation will be > 0.25 in. Setting a cutoff *FP* threshold of 50% for forecasting a "yes", the *FP* = 37% would be a "no" forecast for this event. Although there is actually 37% chance of occurrence, the binary-type forecast is that the event will not occur. If the event does not occur the forecast is called a *correct rejection* (*CR*), but if it does occur the forecast is a *miss* (*M*). Alternatively, a "yes" forecast where the event does occur is termed a *hit* (*H*) and for a non-occurrence, the forecast is a *false alarm* (*FA*).

Varying the cutoff *FP* threshold gives a different false alarm rate (*FAR*) and hit rate (*HR*) which are plotted against each other to produce the ROC (Stanski et al., 1989).



$$HR = \frac{H}{H + M} \quad (42)$$

$$FAR = \frac{FA}{FA + CR} \quad (43)$$

which are built from a contingency table (Figure 66). These two ratios are both concerned with the outcome from a “yes” forecast. The *HR* (hits divided by occurrences) is the fraction of the times when the event did occur that it was forecast to occur. The *FAR* (false alarms divided by nonoccurrences) is the fraction of the times when the event did not occur that it was forecast to occur.

Table 11 uses the same forecast data as Table 10 and shows the values used in calculation of the ROC points of Figure 67. Each *FP* threshold generates a unique contingency table and thus a point in the ROC. A lower *FP* threshold has a high number of hits and a high number of false alarms, thus producing a high *HR* and a high *FAR*. ROC is another way to measure resolution since it reveals the system’s ability to discriminate between occurrences and nonoccurrences, but the ROC does not measure reliability (Evans et al., 2000). A set of forecasts with perfect discrimination has *HR* = 1.0 with a *FAR* = 0.0, so a ROC curve closer to the upper left of the graph represents better forecasts. The diagonal line on the ROC is the zero skill line where forecasts are not able to discriminate at all (Jolliffe and Stephenson, 2003)

The area (*A*) under the ROC curve is an overall measure of the utility of the forecasts from a signal detection point of view. The *A* can be used to produce a ROC skill score (*ROCSS*) akin to the *BSS* (Jolliffe and Stephenson, 2003):

$$ROCSS = 2A - 1 \quad (44)$$

so that a *ROCSS* of 1.0 represents perfect forecasts and a *ROCSS* ≤ 0.0 represents useless forecasts.

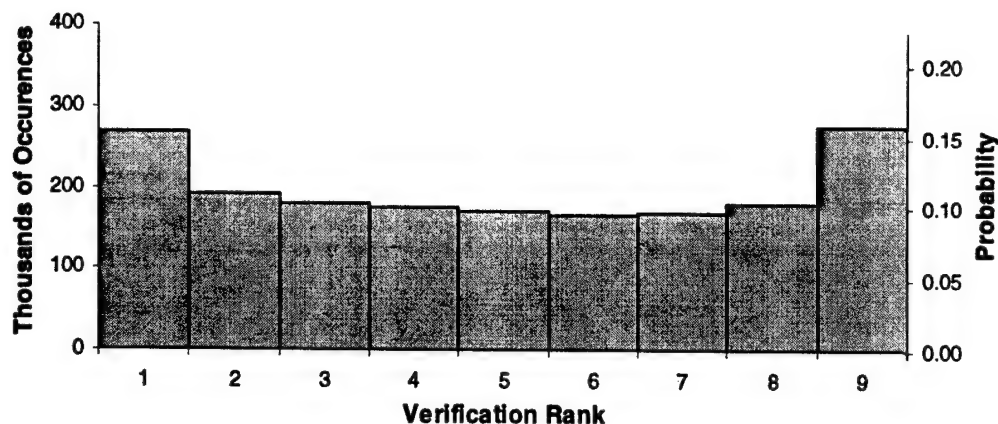


Figure 62. Example VRH for 1,781,676 trials (i.e., forecast/verification data pairs) of  $T_{850}$  using bias-corrected, ACME<sup>core</sup>, 24-h forecasts and centroid analysis verification. The probability for a rank, normally the only quantity displayed, is found by dividing the number of occurrences of the verification in the rank by the total number of trials.

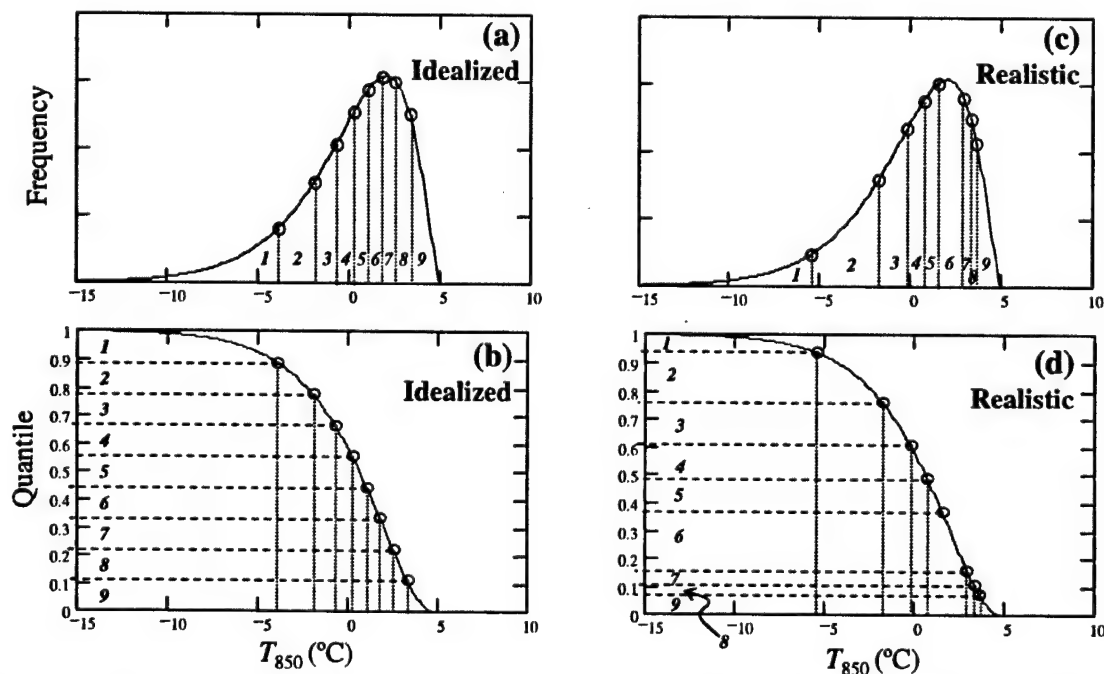


Figure 63. Hypothetical PDF (top) and decumulative density function (bottom) for an idealized EF (a & b) case and realistic EF (c and d) case for forecasts of  $T_{850}$ . Italic numbers label the nine possible ranks in which the verification may occur. Notice that in the idealized case, each bin marks out an equal area under the PDF of exactly  $1/9$ , which corresponds to the equal spacing among the quantiles. In the realistic case, the eight forecasts are random draws that typically do not come out as evenly space quantiles.

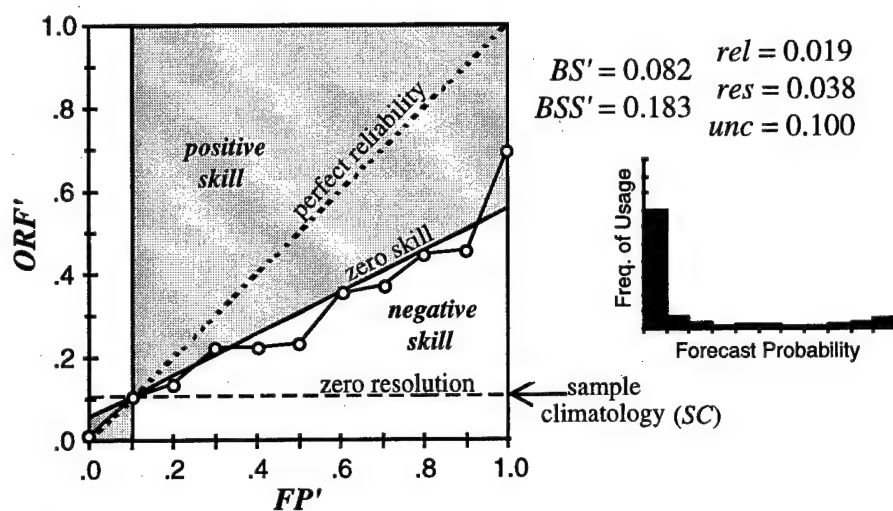


Figure 64. (adapted from Figure 7.8, Wilks, 1995) Reliability diagram for data in Table 10. Open dots (o) showing the observed relative frequency at each tenth of forecast probability are connected with line segments. The shaded area is the *skill zone* in which points make a positive contribution to the  $BSS'$  since  $rel < res$ .

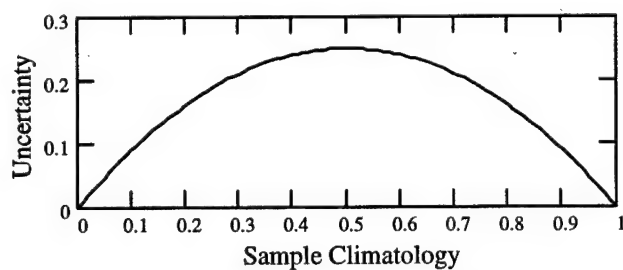


Figure 65. Graph all possible values of the uncertainty term in the  $BS$ . Maximum uncertainty of 0.25 occurs at  $SC = 0.5$ .

Observed	Forecast		
	Yes	No	
	Yes	No	
Yes	$H$	$M$	$H + M = \text{occurrences}$
No	$FA$	$CR$	$FA + CR = \text{non-occurrences}$
	$H + FA$	$M + CR$	Total # of Samples

Figure 66. Contingency table of signal detection theory where  $H$  is number of hits,  $M$  is number of misses,  $FA$  is number of false alarms, and  $CR$  is the number of correct rejections.

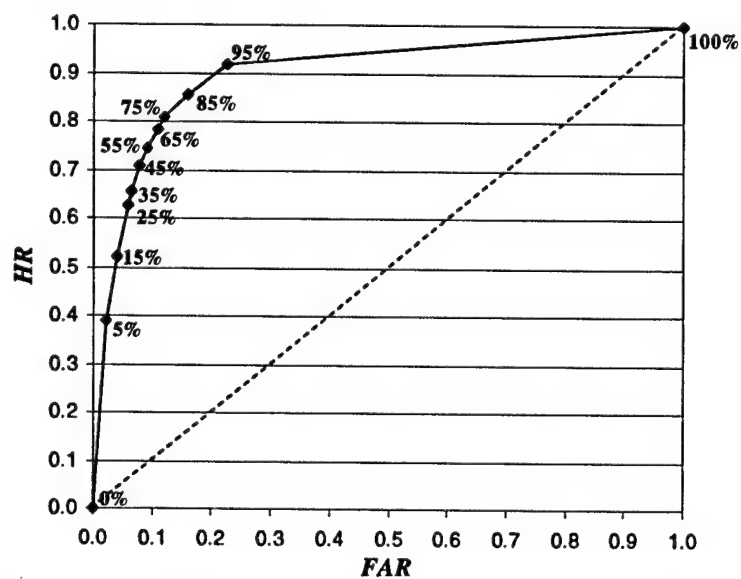


Figure 67. ROC for the data in Table 11. Total area under the curve is 0.90. Points are labeled here (but not normally not included in the ROC) with the  $FP$  threshold that produced each  $FAR$  vs.  $HR$  datum.

Table 9. Two sets of hypothetical EFs of  $T_{850}$  ordered from least to greatest, produced from the PDF in Figure 63. The "idealized forecasts" are evenly spaced quantiles, which only occurs for a long-term average of many realizations. The "realistic forecast" is made up of random draws from the PDF.

EF Member #	Idealized Forecast		Realistic Forecast	
	Quantile	Forecast(°C)	Quantile	Forecast(°C)
1	0.889	-3.93	0.937	-5.47
2	0.778	-1.97	0.761	-1.76
3	0.667	-0.72	0.606	-0.18
4	0.556	0.23	0.490	0.72
5	0.444	1.04	0.370	1.54
6	0.333	1.78	0.161	2.93
7	0.222	2.51	0.108	3.33
8	0.111	3.30	0.072	3.64

Table 10. Summary of 22,402 probability forecasts of 24-h cumulative precipitation > 0.25 inch at the 36-h forecast lead time.  $ORF_i'$  is found by dividing the number of occurrences (Occ.) by the number of forecasts (Fcsts.).

$i$	$FP_i'$	# of Fcsts.	# of Occ.	$ORF_i'$
1	0.0	15609	210	0.01
2	0.1	1483	152	0.10
3	0.2	884	121	0.14
4	0.3	273	61	0.22
5	0.4	457	102	0.22
6	0.5	395	92	0.23
7	0.6	369	130	0.35
8	0.7	209	78	0.37
9	0.8	595	267	0.45
10	0.9	716	328	0.46
11	1.0	1412	990	0.70
TOTALs:		22402	2531	

Table 11. Calculated values for the ROC for the same source data as in Table 10, where the number of non-occurrences is simply the number of forecasts minus the number of occurrences. Each probability threshold in effect produces its own unique contingency table. The arrows give examples of which values are summed to arrive at  $H$ ,  $FA$ ,  $CR$ , and  $M$ .

$FP$ Threshold	# of Occ.	# of Non-occ.	$H$	$FA$	$CR$	$M$	$FAR$	$HR$
0.00	0	0	2321	4472	0	0	1.000	1.000
0.05	210	15399	2321	4472	15399	210	0.225	0.917
0.15	152	1331	2169	3141	16730	362	0.158	0.857
0.25	121	763	2048	2378	17493	483	0.120	0.809
0.35	61	212	1987	2166	17705	544	0.109	0.785
0.45	102	355	1885	1811	18060	646	0.091	0.745
0.55	92	303	1793	1508	18363	738	0.076	0.708
0.65	130	239	1663	1269	18602	868	0.064	0.657
0.75	78	131	1585	1138	18733	946	0.057	0.626
0.85	267	328	1318	810	19061	1213	0.041	0.521
0.95	328	388	990	422	19449	1541	0.021	0.391
1.00	990	422	0	0	19871	2531	0.000	0.000

## Appendix II: ACME<sup>core+</sup> Reference Data

This appendix provides additional material and data on the perturbations to the surface boundary parameters (SBPs) in ACME<sup>core+</sup>.

### A. Uncertainty in Moisture Availability

The goal of this discussion is to demonstrate the difficulty of understanding and quantifying the uncertainty in a model parameterization, a major challenge in designing an EF system. One such parameterization in this research was the MM5 SBP of moisture availability ( $M$ ), for which we endeavored to design a proper perturbation of for ACME<sup>core+</sup>.  $M$  is used in MM5 to model the evaporation rate (i.e., moisture flux at the surface),  $E$ , so that is where this discussion begins.

Determining  $E$  is an essential element of modeling the planetary boundary layer (PBL) since the amount of moisture there greatly determines the evolution of weather phenomena. From Monin-Obhukov similarity theory,  $E$  is described by the bulk transfer relation (Garratt, 1992) where moisture and wind speed are known at measurement height  $h_1$  (e.g., 2 m).

$$\frac{E}{\rho} = (\overline{w'q'})_0 = \frac{q_0 - q_1}{r_a} \quad (45)$$

$$r_a = 1/C_H V_1$$

where  $\rho$  is the standard atmosphere's surface air density ( $1.23 \text{ kg m}^{-3}$ ),  $(\overline{w'q'})_0$  is moisture flux at the surface [ $\text{m s}^{-1}$ ],  $q_0$  is the mixing ratio at the surface [ ],  $q_1$  is the mixing ratio at  $h_1$  [ ],  $r_a$  is the aerodynamic resistance [ $\text{s m}^{-1}$ ],  $C_H$  is the drag coefficient (function of  $V_1$ ,  $h_1$ , surface roughness length  $z_0$ , thermal roughness length  $z_T$ , and  $L$ ) [ ], and  $V_1$  is the wind speed at  $h_1$  [ $\text{m s}^{-1}$ ].

Evaporation is the mass of water vapor leaving a unit surface area, per unit time, that is replacing water vapor fluxing up toward air with a lower mixing ratio ( $q_1$ ). Note that if  $q_1 > q_0$  then

$(\overline{w'q'})_0$  is reversed and dew or frost forms at the surface. While this equation works well over water, it is an overestimate over land where  $E$  can be limited by unsaturated soil and vegetation.

A more thorough description of evaporation is obtained by including the Clausius Clapeyron equation, energy balance, and stomatal resistance. This gives an equation for the evaporation over an ideal (i.e., saturated), vegetated surface (Bretherton, 2002).

$$E = \frac{\Gamma^*(R_N - H_G)}{L} + \frac{(1 - \Gamma^*)\rho(q_1^* - q_1)}{r_{st} + r_a} \quad (46)$$

$$\Gamma^* = \frac{s^*}{s^* + 1 + \frac{r_{st}}{r_a}} \quad s^* = \frac{L}{C_P} \left( \frac{dq^*}{dT} \right)_{T_R} \quad r_{st} = \frac{\rho(q_0^* - q_0)}{E}$$

where  $R_N$  is the net downward radiative flux at the surface [ $\text{W m}^{-2}$ ],  $H_G$  is the downward ground heat flux [ $\text{W m}^{-2}$ ],  $L$  is the latent heat of evaporation ( $2.5 \times 10^6 \text{ J kg}^{-1}$ ),  $C_P$  is the specific heat at constant pressure ( $1004 \text{ J K}^{-1} \text{ kg}^{-1}$ ),  $T_R$  is a reference temperature [K],  $q_1^*$  is the saturation mixing ratio at  $h_1$  (function of  $T_1$ ) [ ],  $q_0^*$  is the saturation mixing ratio at the surface (function of  $T_0$ ) [ ], and  $q_0$  is the mixing ratio at the surface [ ].

For a nonvegetated surface that is not saturated, the evaporation is also a function of available soil moisture, described by the surface relative humidity,  $RH_0$ .

$$E = \frac{\rho}{r_a} (RH_0 q_0^* - q_1) \quad (47)$$

Ideally then, a mesoscale model would use Equation (46) in a wet, vegetated grid box, Equation (47) over a nonvegetated soil or water grid box, and some appropriate amalgamation of the two for a mixed grid box. While such a strategy is possible, it is not practical for a real-time modeling system where parameterizations are designed to give reasonable results without being too demanding of processing power. A big concern for a real-time mesoscale model is completing



the simulation quickly in order to make products available to weather forecasters in a timely manner.

The MM5 models  $E$  by using  $M$  to represent the effects of stomatal resistance, aerodynamic resistance, and soil moisture.

$$E = M C_H V_1 \rho (q_0^* - q_1) \quad (48)$$

which most closely resembles Equation (47) except  $r_a$  is now replaced by the total resistance,  $r = r_a + r_{st}$ . This means that MM5 considers

$$r = \frac{1}{M C_H V_1} \Rightarrow r_a + r_{st} = \frac{1}{M} r_a \Rightarrow r_{st} = r_a \left( \frac{1}{M} - 1 \right) \quad (49)$$

revealing the primary fault of this parameterization scheme. The stomatal resistance should not be modeled as a function of the aerodynamic resistance since  $r_a$  and  $r_{st}$  are independent. Additionally, it is difficult to see how  $M$  could possibly represent soil moisture at the same time. Further uncertainty in  $M$  is also provided by the use of the land use table, as discussed in the main text.

While this discussion certainly provides an understanding of the uncertainty in  $M$ , it is clear that there is no practical way to quantify the uncertainty. The problem is the same with other SBPs and can get even worse when considering other model aspects. This is the chief challenge in designing an EF system such as ACME<sup>core+</sup> that attempts to account for model uncertainty with model perturbations.

## B. Land Use Table

Table 12 –Table 14 provide the values of the gamma variables applied to Equation (25) to generate perturbed values of the surface boundary parameters (SBPs). Table 15 then shows the eight perturbed land use tables, which were generated using random deviates from the SBP PDFs.

Figure 68 provides a plot of the fixed SST perturbation field for each of the 8 members of

ACME<sup>core+</sup>. These perturbations were applied to the daily OTIS SST field (used by the ACME<sup>core</sup> members) to produce a unique SST for each ACME<sup>core+</sup> member.

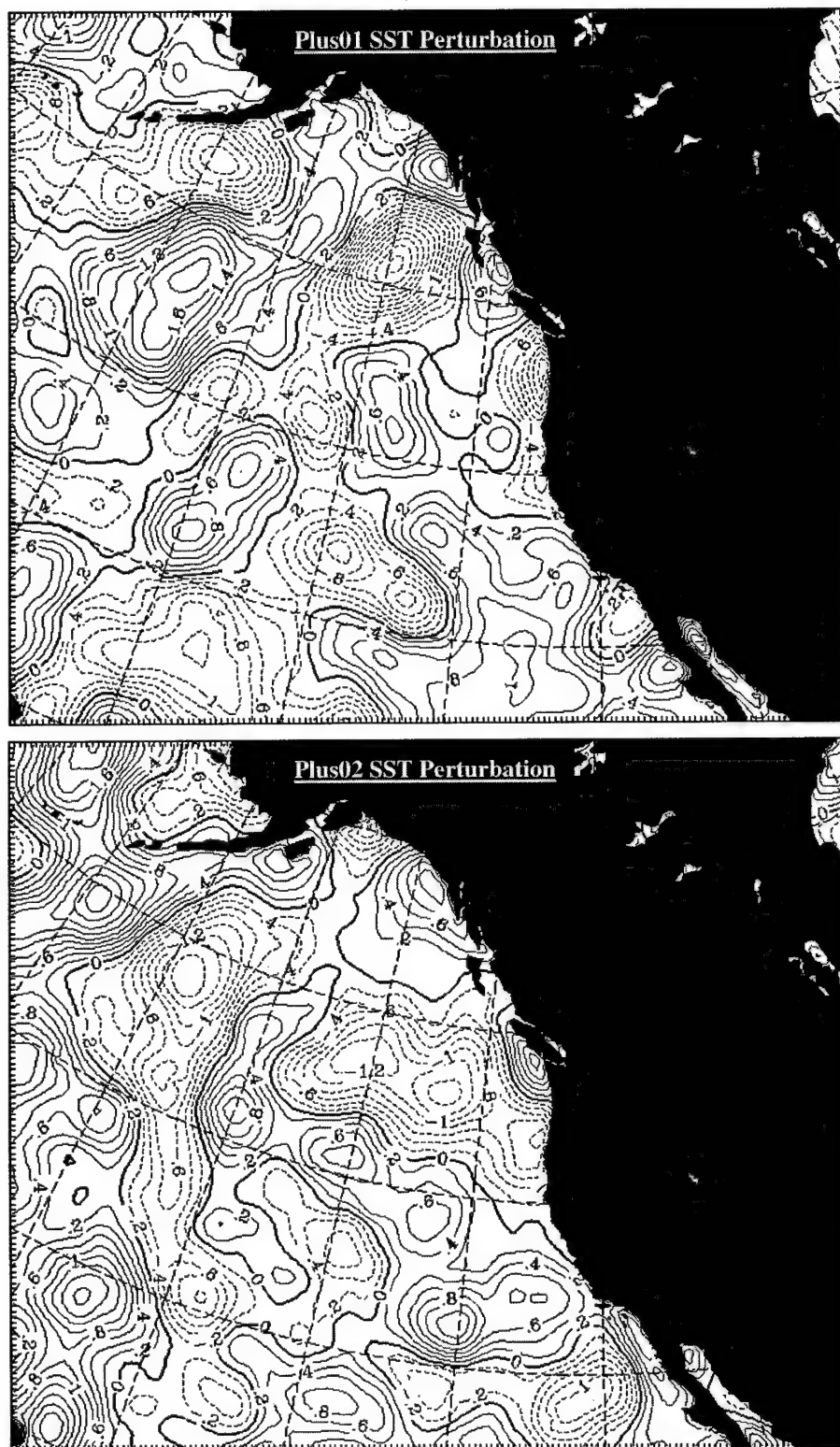


Figure 68. Plots of the SST perturbations for each member of ACME<sup>core+</sup>.

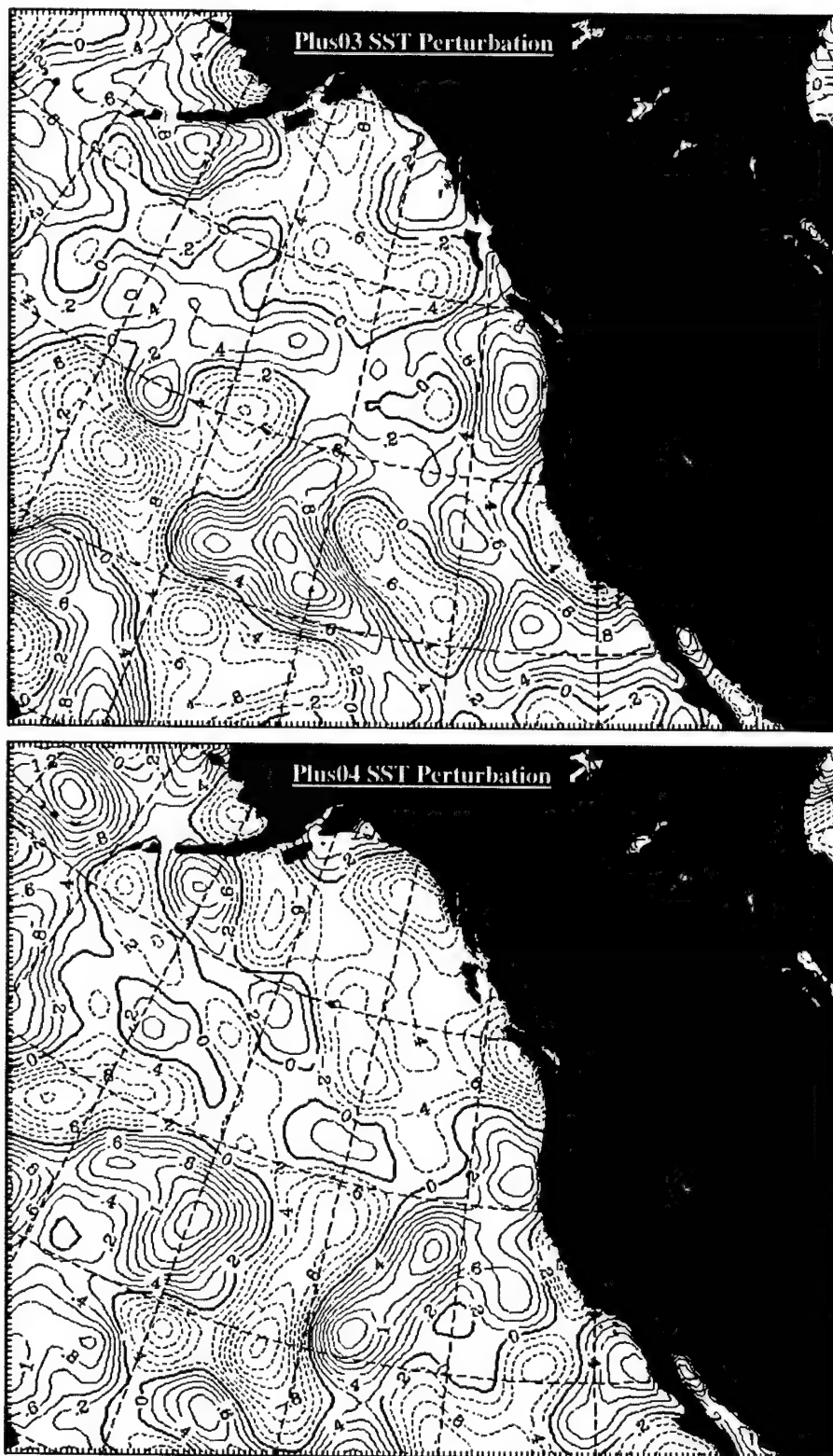


Figure 68. Continued

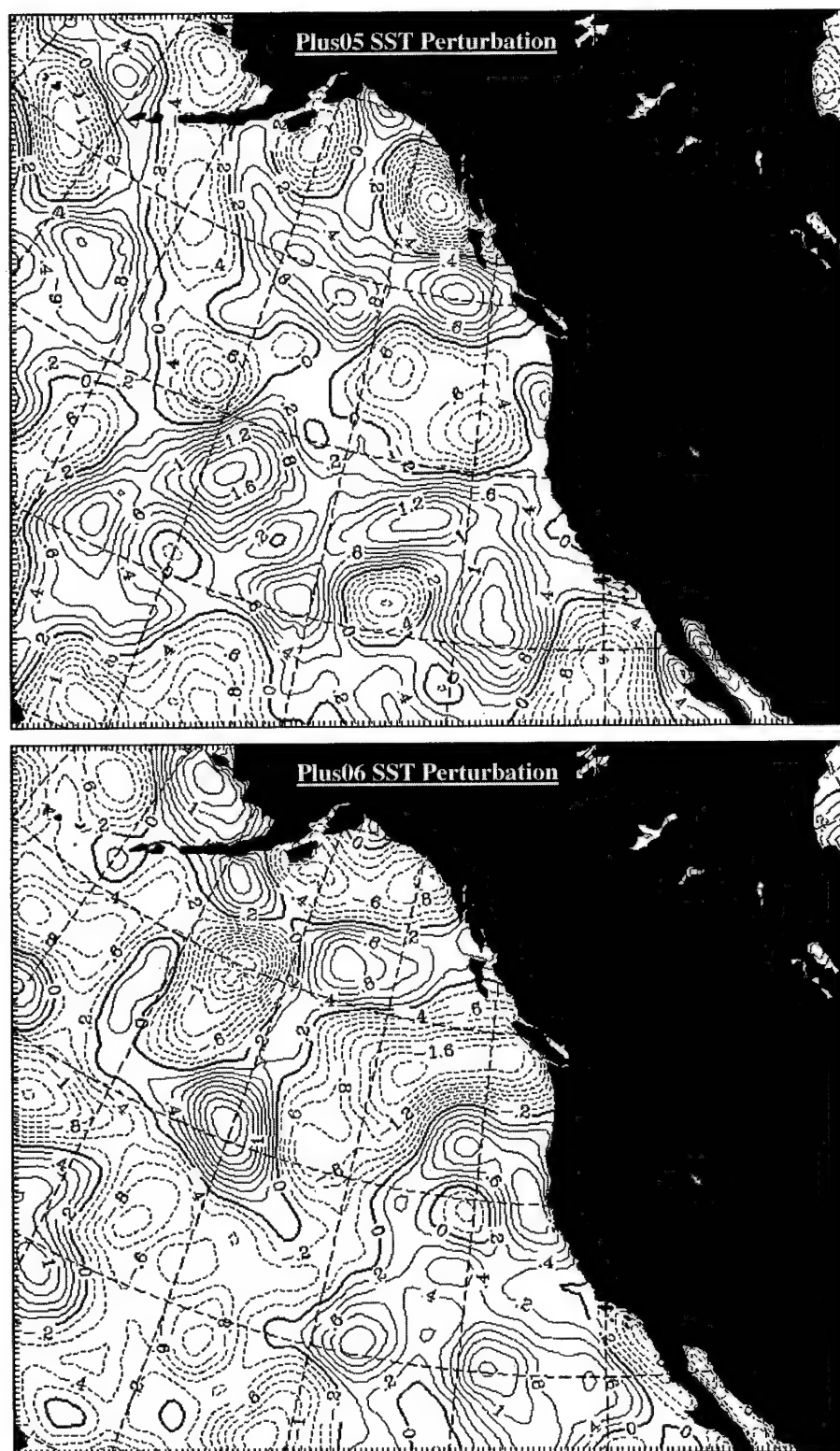


Figure 68. Continued

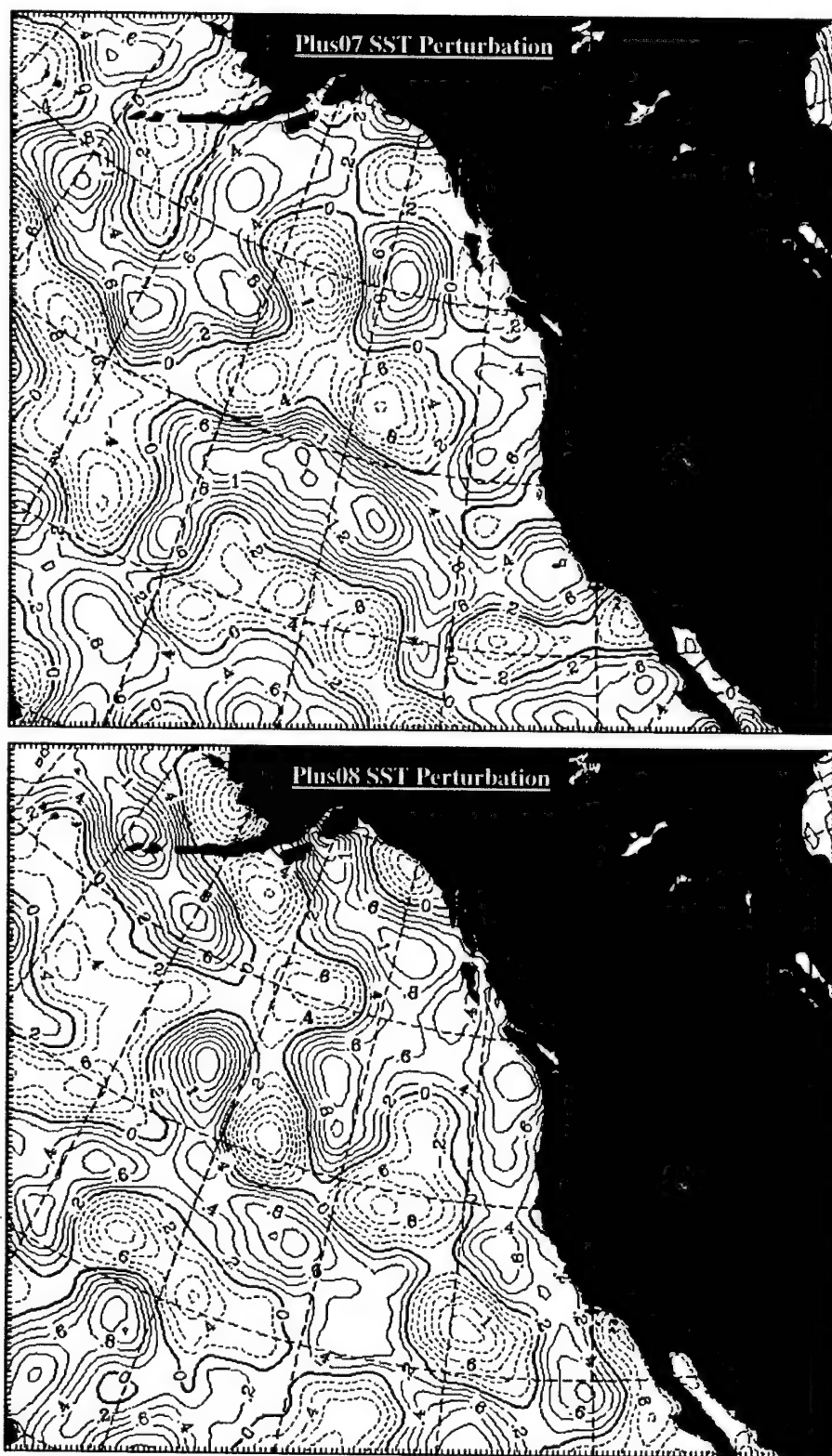


Figure 68. Continued

Table 12. Gamma variables for the 48 PDFs used to generate albedo values.

Land Use #	<u>Albedo</u>							
	Summer				Winter			
	reverse	shape	spread	translate	reverse	shape	spread	translate
1	1	50	0.57	9.85	1	50	0.566	9.85
2	1	50	0.57	10.85	1	50	0.566	4.85
3	1	50	0.57	9.85	1	50	0.566	4.85
4	1	50	0.57	9.85	1	50	0.566	4.85
5	1	50	0.57	9.85	1	50	0.566	4.85
6	1	50	0.57	11.85	1	50	0.566	7.85
7	1	50	0.57	8.85	1	50	0.566	4.85
8	1	50	0.57	5.85	1	50	0.566	2.85
9	1	50	0.57	7.85	1	50	0.566	3.85
10	1	100	0.35	14.64	1	100	0.35	14.64
11	1	100	0.30	13.67	1	100	0.3	12.67
12	1	100	0.30	15.67	1	100	0.3	14.67
13	1	100	0.25	12.80	1	100	0.25	12.8
14	1	100	0.25	12.80	1	100	0.25	12.8
15	1	10	1.11	-3.04	1	10	1.11	-4.04
16	-1	30	0.31	-17.00	-1	30	0.31	-17
17	1	100	0.35	20.66	1	100	0.35	20.66
18	1	100	0.35	20.66	1	100	0.35	20.66
19	1	100	0.30	4.72	1	100	0.3	4.72
20	1	3	4.62	-5.83	-1	3	4.62	-69.18
21	1	3	4.04	-6.96	-1	3	4.04	-58.05
22	1	3	4.33	-6.39	-1	3	4.33	-63.62
23	1	3	4.62	-15.83	-1	3	4.62	-79.18
24	-1	10	2.21	-74.81	1	15	1.81	-44.74

Table 13. Gamma variables for the 48 PDFs used to generate moisture availability values.

Land Use #	<u>Moisture Availability</u>							
	Summer				Winter			
	reverse	shape	spread	translate	reverse	shape	spread	translate
1	1	4	3.00	-0.98	1	4	3	-0.98
2	1	3.5	5.88	-15.22	-1	5	4.92	-79.81
3	1	500	0.45	173.20	1	500	0.447	173.2
4	1	3.5	5.35	-11.59	-1	7	3.4	-70.28
5	1	3.5	3.47	-16.24	-1	7	2.46	-54.73
6	1	3.5	5.35	-21.59	-1	5	4.47	-77.98
7	1	3	3.18	-8.58	-1	10	1.9	-47.08
8	1	3.5	2.14	-4.66	-1	8	1.24	-28.63
9	1	3.5	2.51	-8.77	-1	9	1.57	-37.52
10	1	4	3.00	-5.98	1	4	3	-5.98
11	1	3.5	5.88	-15.22	-1	5	4.92	-79.81
12	1	3.5	5.88	-15.22	-1	5	4.92	-79.81
13	1	500	0.45	173.20	1	500	0.447	173.2
14	1	3.5	5.88	-15.22	-1	5	4.92	-79.81
15	1	3.5	5.88	-15.22	-1	5	4.92	-79.81
16	1	99	99.00	99.00	1	99	99	99
17	1	6	2.65	-46.78	-1	4.5	3.11	-85.74
18	1	3	6.35	-22.05	-1	3	6.35	-82.93
19	1	2	1.06	-0.94	-1	10	0.32	-7.85
20	1	3.5	5.08	-37.09	-1	2	7.78	-97.85
21	1	3.5	5.08	-37.09	-1	2	7.78	-97.85
22	1	3.5	5.08	-37.09	-1	2	7.78	-97.85
23	1	1.1	10.49	-0.95	-1	1.1	10.49	-96.04
24	-1	3	1.27	-97.54	-1	3	1.27	-97.54

Table 14. Gamma variables for the 48 PDFs used to generate roughness length values.

Land Use #	<b>Roughness Length</b> Summer				Winter			
	reverse	shape	spread	translate	reverse	shape	spread	translate
1	1	2	21.21	-28.86	1	2	21.21	-28.86
2	1	100	0.35	19.66	1	3	2.02	-0.95
3	1	100	0.35	19.66	1	3	2.02	-0.95
4	1	100	0.35	19.66	1	3	2.02	-0.95
5	1	100	0.35	20.66	1	3	2.02	-0.95
6	1	100	0.50	29.50	1	100	0.5	29.5
7	-1	30	0.50	-26.30	-1	30	0.46	-23.24
8	1	100	0.25	14.78	1	100	0.25	14.78
9	1	100	0.25	13.78	1	100	0.25	14.78
10	1	100	0.25	9.78	1	100	0.25	9.78
11	1	1.5	34.30	-32.52	1	1.5	34.3	-32.52
12	1	1.5	34.30	-32.52	1	1.5	34.3	-32.52
13	1	1.5	34.30	-32.52	1	1.5	34.3	-32.52
14	1	1.5	34.30	-32.52	1	1.5	34.3	-32.52
15	1	1.5	34.30	-32.52	1	1.5	34.3	-32.52
16	1	1.3	0.02	0.00	1	1.3	0.0175	-0.0047
17	1	100	0.50	29.50	1	100	0.5	29.5
18	1	3	6.35	-27.45	1	3	6.35	-27.45
19	1	100	0.25	14.78	1	100	0.25	14.78
20	1	100	0.25	14.78	1	100	0.25	14.78
21	1	3	6.35	-17.45	1	3	6.35	-17.45
22	1	10	1.26	-3.64	1	10	1.26	-3.64
23	1	10	1.11	-0.04	1	3	2.02	-0.95
24	-1	15	0.28	-8.97	-1	15	0.28	-8.97



Table 15. MM5 land use tables used for ACME<sup>core+</sup>. The format shown is exactly as a land use file is employed in the MM5 code. The perturbed parameters are albedo (ALBD, %), moisture availability (SLMO, % \* 100), and roughness length (SFZO, cm). Parameters not perturbed include emissivity (SFEM, % \* 100), thermal inertia (THERIN), snow-effect factor (SCFX), and heat capacity (SFHC).

LANDUSE.TBL.plus01

USGS 24,2,	ALBD	SLMO	SFEM	SFZO	THERIN	SCFX	SFHC	
SUMMER								
1,	16.,	.16,	.88,	49.,	3.,	.52,	18.9e5,	'Urban and Built-Up Land'
2,	17.,	.37,	.92,	20.,	4.,	.60,	25.0e5,	'Dryland Cropland and Pasture'
3,	20.,	.33,	.92,	12.,	4.,	.60,	25.0e5,	'Irrigated Cropland and Pasture'
4,	25.,	.32,	.92,	9.,	4.,	.60,	25.0e5,	'Mixed Dryland/Irrigated Cropland and Pasture'
5,	25.,	.28,	.92,	12.,	4.,	.60,	25.0e5,	'Cropland/Grassland Mosaic'
6,	14.,	.84,	.93,	29.,	4.,	.60,	25.0e5,	'Cropland/Woodland Mosaic'
7,	16.,	.35,	.92,	17.,	3.,	.60,	20.8e5,	'Grassland'
8,	20.,	.09,	.88,	6.,	3.,	.62,	20.8e5,	'Shrubland'
9,	14.,	.15,	.90,	9.,	3.,	.60,	20.8e5,	'Mixed Shrubland/Grassland'
10,	18.,	.24,	.92,	20.,	3.,	0.,	25.0e5,	'Savanna'
11,	16.,	.43,	.93,	51.,	4.,	.56,	25.0e5,	'Deciduous Broadleaf Forest'
12,	17.,	.68,	.94,	94.,	4.,	.50,	25.0e5,	'Deciduous Needleleaf Forest'
13,	10.,	.46,	.95,	145.,	5.,	0.,	29.2e5,	'Evergreen Broadleaf Forest'
14,	10.,	.60,	.95,	95.,	4.,	.50,	29.2e5,	'Evergreen Needleleaf Forest'
15,	13.,	.56,	.94,	123.,	4.,	.54,	41.8e5,	'Mixed Forest'
16,	8.,	1.0,	.98,	0.01,	6.,	0.,	9.0e25,	'Water Bodies'
17,	14.,	.63,	.95,	23.,	6.,	.55,	29.2e5,	'Herbaceous Wetland'
18,	19.,	.34,	.95,	41.,	5.,	.58,	41.8e5,	'Wooded Wetland'
19,	23.,	.03,	.85,	15.,	2.,	.62,	12.0e5,	'Barren or Sparsely Vegetated'
20,	14.,	.49,	.92,	14.,	5.,	.60,	9.0e25,	'Herbaceous Tundra'
21,	38.,	.59,	.93,	37.,	5.,	.60,	9.0e25,	'Wooded Tundra'
22,	24.,	.43,	.92,	16.,	5.,	.60,	9.0e25,	'Mixed Tundra'
23,	35.,	.08,	.85,	10.,	2.,	.62,	12.0e5,	'Bare Ground Tundra'
24,	57.,	.96,	.95,	6.,	5.,	0.,	9.0e25,	'Snow or Ice'
WINTER								
1,	13.,	.13,	.88,	54.,	3.,	.52,	18.9e5,	'Urban and Built-Up Land'
2,	25.,	.32,	.92,	6.,	4.,	.60,	25.0e5,	'Dryland Cropland and Pasture'
3,	20.,	.66,	.92,	4.,	4.,	.60,	25.0e5,	'Irrigated Cropland and Pasture'
4,	15.,	.43,	.92,	3.,	4.,	.60,	25.0e5,	'Mixed Dryland/Irrigated Cropland and Pasture'
5,	25.,	.24,	.92,	7.,	4.,	.60,	25.0e5,	'Cropland/Grassland Mosaic'
6,	16.,	.65,	.93,	21.,	4.,	.60,	25.0e5,	'Cropland/Woodland Mosaic'
7,	23.,	.27,	.92,	9.,	4.,	.60,	20.8e5,	'Grassland'
8,	18.,	.12,	.88,	12.,	4.,	.62,	20.8e5,	'Shrubland'
9,	19.,	.23,	.90,	8.,	4.,	.60,	20.8e5,	'Mixed Shrubland/Grassland'
10,	13.,	.19,	.92,	9.,	3.,	0.,	25.0e5,	'Savanna'
11,	20.,	.15,	.93,	65.,	5.,	.56,	25.0e5,	'Deciduous Broadleaf Forest'
12,	16.,	.60,	.93,	77.,	5.,	.50,	25.0e5,	'Deciduous Needleleaf Forest'
13,	12.,	.66,	.95,	80.,	5.,	0.,	29.2e5,	'Evergreen Broadleaf Forest'
14,	15.,	.33,	.95,	44.,	5.,	.50,	29.2e5,	'Evergreen Needleleaf Forest'
15,	16.,	.64,	.94,	56.,	6.,	.58,	41.8e5,	'Mixed Forest'
16,	8.,	1.0,	.98,	0.01,	6.,	0.,	9.0e25,	'Water Bodies'
17,	15.,	.70,	.95,	26.,	6.,	.55,	29.2e5,	'Herbaceous Wetland'
18,	12.,	.71,	.95,	49.,	6.,	.58,	41.8e5,	'Wooded Wetland'
19,	24.,	.03,	.85,	7.,	2.,	.62,	12.0e5,	'Barren or Sparsely Vegetated'
20,	56.,	.83,	.92,	9.,	5.,	0.,	9.0e25,	'Herbaceous Tundra'
21,	51.,	.61,	.93,	20.,	5.,	0.,	9.0e25,	'Wooded Tundra'
22,	49.,	.67,	.92,	16.,	5.,	0.,	9.0e25,	'Mixed Tundra'
23,	59.,	.85,	.95,	6.,	5.,	0.,	12.0e5,	'Bare Ground Tundra'
24,	79.,	.93,	.95,	5.,	5.,	0.,	9.0e25,	'Snow or Ice'

Table 15 continued:

LANDUSE.TBL.plus02

USGS									
24,2	'ALBD	SLMO	SFEM	SFZO	THERIN	SCFX	SFHC		
SUMMER									
1,	16.,	.10,	.88,	51.,	3.,	.52,	18.9e5,	'Urban and Built-Up Land'	
2,	23.,	.46,	.92,	11.,	4.,	.60,	25.0e5,	'Dryland Cropland and Pasture'	
3,	16.,	.52,	.92,	18.,	4.,	.60,	25.0e5,	'Irrigated Cropland and Pasture'	
4,	16.,	.28,	.92,	18.,	4.,	.60,	25.0e5,	'Mixed Dryland/Irrigated Cropland and Pasture'	
5,	17.,	.32,	.92,	13.,	4.,	.60,	25.0e5,	'Cropland/Grassland Mosaic'	
6,	18.,	.54,	.93,	17.,	4.,	.60,	25.0e5,	'Cropland/Woodland Mosaic'	
7,	19.,	.13,	.92,	13.,	3.,	.60,	20.8e5,	'Grassland'	
8,	19.,	.12,	.88,	13.,	3.,	.62,	20.8e5,	'Shrubland'	
9,	24.,	.18,	.90,	12.,	3.,	.60,	20.8e5,	'Mixed Shrubland/Grassland'	
10,	26.,	.31,	.92,	12.,	3.,	0.,	25.0e5,	'Savanna'	
11,	20.,	.19,	.93,	54.,	4.,	.56,	25.0e5,	'Deciduous Broadleaf Forest'	
12,	16.,	.38,	.94,	64.,	4.,	.50,	25.0e5,	'Deciduous Needleleaf Forest'	
13,	13.,	.51,	.95,	111.,	5.,	0.,	29.2e5,	'Evergreen Broadleaf Forest'	
14,	15.,	.20,	.95,	66.,	4.,	.50,	29.2e5,	'Evergreen Needleleaf Forest'	
15,	17.,	.30,	.94,	85.,	4.,	.54,	41.8e5,	'Mixed Forest'	
16,	8.,	1.0,	.98,	0.01,	6.,	0.,	9.0e25,	'Water Bodies'	
17,	14.,	.60,	.95,	26.,	6.,	.55,	29.2e5,	'Herbaceous Wetland'	
18,	14.,	.53,	.95,	52.,	5.,	.58,	41.8e5,	'Wooded Wetland'	
19,	21.,	.02,	.85,	12.,	2.,	.62,	12.0e5,	'Barren or Sparsely Vegetated'	
20,	14.,	.41,	.92,	9.,	5.,	.60,	9.0e25,	'Herbaceous Tundra'	
21,	13.,	.64,	.93,	33.,	5.,	.60,	9.0e25,	'Wooded Tundra'	
22,	12.,	.50,	.92,	17.,	5.,	.60,	9.0e25,	'Mixed Tundra'	
23,	27.,	.07,	.85,	10.,	2.,	.62,	12.0e5,	'Bare Ground Tundra'	
24,	49.,	.85,	.95,	5.,	5.,	0.,	9.0e25,	'Snow or Ice'	
WINTER									
1,	22.,	.14,	.88,	46.,	3.,	.52,	18.9e5,	'Urban and Built-Up Land'	
2,	26.,	.16,	.92,	4.,	4.,	.60,	25.0e5,	'Dryland Cropland and Pasture'	
3,	18.,	.60,	.92,	9.,	4.,	.60,	25.0e5,	'Irrigated Cropland and Pasture'	
4,	22.,	.56,	.92,	7.,	4.,	.60,	25.0e5,	'Mixed Dryland/Irrigated Cropland and Pasture'	
5,	22.,	.33,	.92,	2.,	4.,	.60,	25.0e5,	'Cropland/Grassland Mosaic'	
6,	17.,	.65,	.93,	31.,	4.,	.60,	25.0e5,	'Cropland/Woodland Mosaic'	
7,	29.,	.30,	.92,	11.,	4.,	.60,	20.8e5,	'Grassland'	
8,	23.,	.16,	.88,	15.,	4.,	.62,	20.8e5,	'Shrubland'	
9,	23.,	.29,	.90,	12.,	4.,	.60,	20.8e5,	'Mixed Shrubland/Grassland'	
10,	18.,	.15,	.92,	16.,	3.,	0.,	25.0e5,	'Savanna'	
11,	17.,	.63,	.93,	57.,	5.,	.56,	25.0e5,	'Deciduous Broadleaf Forest'	
12,	17.,	.72,	.93,	66.,	5.,	.50,	25.0e5,	'Deciduous Needleleaf Forest'	
13,	16.,	.58,	.95,	138.,	5.,	0.,	29.2e5,	'Evergreen Broadleaf Forest'	
14,	12.,	.67,	.95,	74.,	5.,	.50,	29.2e5,	'Evergreen Needleleaf Forest'	
15,	11.,	.64,	.94,	54.,	6.,	.58,	41.8e5,	'Mixed Forest'	
16,	8.,	1.0,	.98,	0.01,	6.,	0.,	9.0e25,	'Water Bodies'	
17,	17.,	.79,	.95,	20.,	6.,	.55,	29.2e5,	'Herbaceous Wetland'	
18,	19.,	.75,	.95,	58.,	6.,	.58,	41.8e5,	'Wooded Wetland'	
19,	25.,	.04,	.85,	9.,	2.,	.62,	12.0e5,	'Barren or Sparsely Vegetated'	
20,	68.,	.93,	.92,	16.,	5.,	0.,	9.0e25,	'Herbaceous Tundra'	
21,	47.,	.39,	.93,	48.,	5.,	0.,	9.0e25,	'Wooded Tundra'	
22,	55.,	.96,	.92,	22.,	5.,	0.,	9.0e25,	'Mixed Tundra'	
23,	71.,	.94,	.95,	7.,	5.,	0.,	12.0e5,	'Bare Ground Tundra'	
24,	70.,	.93,	.95,	5.,	5.,	0.,	9.0e25,	'Snow or Ice'	

Table 15 continued:

LANDUSE.TBL.plus03

USGS	24,2	'ALBD	SLMO	SPFM	SFZO	THERIN	SCFX	SFHC	
SUMMER									
1,	20.,	.06,	.88,	98.,	3.,	.52,	18.9e5,	'Urban and Built-Up Land'	
2,	13.,	.37,	.92,	14.,	4.,	.60,	25.0e5,	'Dryland Cropland and Pasture'	
3,	23.,	.56,	.92,	18.,	4.,	.60,	25.0e5,	'Irrigated Cropland and Pasture'	
4,	20.,	.25,	.92,	12.,	4.,	.60,	25.0e5,	'Mixed Dryland/Irrigated Cropland and Pasture'	
5,	15.,	.35,	.92,	17.,	4.,	.60,	25.0e5,	'Cropland/Grassland Mosaic'	
6,	18.,	.48,	.93,	32.,	4.,	.60,	25.0e5,	'Cropland/Woodland Mosaic'	
7,	20.,	.13,	.92,	9.,	3.,	.60,	20.8e5,	'Grassland'	
8,	33.,	.10,	.88,	10.,	3.,	.62,	20.8e5,	'Shrubland'	
9,	18.,	.23,	.90,	4.,	3.,	.60,	20.8e5,	'Mixed Shrubland/Grassland'	
10,	22.,	.35,	.92,	13.,	3.,	0.,	25.0e5,	'Savanna'	
11,	17.,	.38,	.93,	205.,	4.,	.56,	25.0e5,	'Deciduous Broadleaf Forest'	
12,	11.,	.31,	.94,	108.,	4.,	.50,	25.0e5,	'Deciduous Needleleaf Forest'	
13,	10.,	.47,	.95,	69.,	5.,	0.,	29.2e5,	'Evergreen Broadleaf Forest'	
14,	11.,	.41,	.95,	44.,	4.,	.50,	29.2e5,	'Evergreen Needleleaf Forest'	
15,	19.,	.26,	.94,	129.,	4.,	.54,	41.8e5,	'Mixed Forest'	
16,	8.,	1.0,	.98,	0.01,	6.,	0.,	9.0e25,	'Water Bodies'	
17,	16.,	.58,	.95,	19.,	6.,	.55,	29.2e5,	'Herbaceous Wetland'	
18,	13.,	.49,	.95,	44.,	5.,	.58,	41.8e5,	'Wooded Wetland'	
19,	27.,	.08,	.85,	11.,	2.,	.62,	12.0e5,	'Barren or Sparsely Vegetated'	
20,	24.,	.56,	.92,	10.,	5.,	.60,	9.0e25,	'Herbaceous Tundra'	
21,	22.,	.51,	.93,	36.,	5.,	.60,	9.0e25,	'Wooded Tundra'	
22,	16.,	.58,	.92,	11.,	5.,	.60,	9.0e25,	'Mixed Tundra'	
23,	29.,	.41,	.85,	13.,	2.,	.62,	12.0e5,	'Bare Ground Tundra'	
24,	56.,	.92,	.95,	3.,	5.,	0.,	9.0e25,	'Snow or Ice'	
WINTER									
1,	20.,	.21,	.88,	79.,	3.,	.52,	18.9e5,	'Urban and Built-Up Land'	
2,	23.,	.38,	.92,	5.,	4.,	.60,	25.0e5,	'Dryland Cropland and Pasture'	
3,	27.,	.46,	.92,	4.,	4.,	.60,	25.0e5,	'Irrigated Cropland and Pasture'	
4,	19.,	.63,	.92,	9.,	4.,	.60,	25.0e5,	'Mixed Dryland/Irrigated Cropland and Pasture'	
5,	24.,	.41,	.92,	7.,	4.,	.60,	25.0e5,	'Cropland/Grassland Mosaic'	
6,	24.,	.53,	.93,	24.,	4.,	.60,	25.0e5,	'Cropland/Woodland Mosaic'	
7,	26.,	.21,	.92,	7.,	4.,	.60,	20.8e5,	'Grassland'	
8,	27.,	.17,	.88,	14.,	4.,	.62,	20.8e5,	'Shrubland'	
9,	35.,	.21,	.90,	10.,	4.,	.60,	20.8e5,	'Mixed Shrubland/Grassland'	
10,	13.,	.16,	.92,	17.,	3.,	0.,	25.0e5,	'Savanna'	
11,	21.,	.55,	.93,	48.,	5.,	.56,	25.0e5,	'Deciduous Broadleaf Forest'	
12,	15.,	.31,	.93,	63.,	5.,	.50,	25.0e5,	'Deciduous Needleleaf Forest'	
13,	9.,	.52,	.95,	47.,	5.,	0.,	29.2e5,	'Evergreen Broadleaf Forest'	
14,	18.,	.56,	.95,	81.,	5.,	.50,	29.2e5,	'Evergreen Needleleaf Forest'	
15,	16.,	.58,	.94,	41.,	6.,	.58,	41.8e5,	'Mixed Forest'	
16,	8.,	1.0,	.98,	0.01,	6.,	0.,	9.0e25,	'Water Bodies'	
17,	13.,	.70,	.95,	19.,	6.,	.55,	29.2e5,	'Herbaceous Wetland'	
18,	22.,	.66,	.95,	34.,	6.,	.58,	41.8e5,	'Wooded Wetland'	
19,	24.,	.06,	.85,	6.,	2.,	.62,	12.0e5,	'Barren or Sparsely Vegetated'	
20,	49.,	.91,	.92,	10.,	5.,	0.,	9.0e25,	'Herbaceous Tundra'	
21,	55.,	.70,	.93,	23.,	5.,	0.,	9.0e25,	'Wooded Tundra'	
22,	55.,	.61,	.92,	26.,	5.,	0.,	9.0e25,	'Mixed Tundra'	
23,	71.,	.69,	.95,	3.,	5.,	0.,	12.0e5,	'Bare Ground Tundra'	
24,	85.,	.95,	.95,	7.,	5.,	0.,	9.0e25,	'Snow or Ice'	

Table 15 continued:

LANDUSE.TBL.plus04

USGS	24,2	'ALBD	SLMO	SFEM	SFZ0	THERIN	SCFX	SFHC	
SUMMER									
1,	23.,	.21,	.88,	117.,	3.,	.52,	18.9e5,	'Urban and Built-Up Land'	
2,	13.,	.24,	.92,	15.,	4.,	.60,	25.0e5,	'Dryland Cropland and Pasture'	
3,	15.,	.60,	.92,	13.,	4.,	.60,	25.0e5,	'Irrigated Cropland and Pasture'	
4,	25.,	.28,	.92,	17.,	4.,	.60,	25.0e5,	'Mixed Dryland/Irrigated Cropland and Pasture'	
5,	18.,	.28,	.92,	8.,	4.,	.60,	25.0e5,	'Cropland/Grassland Mosaic'	
6,	24.,	.37,	.93,	26.,	4.,	.60,	25.0e5,	'Cropland/Woodland Mosaic'	
7,	18.,	.16,	.92,	15.,	3.,	.60,	20.8e5,	'Grassland'	
8,	23.,	.12,	.88,	4.,	3.,	.62,	20.8e5,	'Shrubland'	
9,	20.,	.19,	.90,	9.,	3.,	.60,	20.8e5,	'Mixed Shrubland/Grassland'	
10,	18.,	.09,	.92,	16.,	3.,	0.,	25.0e5,	'Savanna'	
11,	14.,	.57,	.93,	284.,	4.,	.56,	25.0e5,	'Deciduous Broadleaf Forest'	
12,	15.,	.22,	.94,	39.,	4.,	.50,	25.0e5,	'Deciduous Needleleaf Forest'	
13,	8.,	.39,	.95,	74.,	5.,	0.,	29.2e5,	'Evergreen Broadleaf Forest'	
14,	15.,	.31,	.95,	175.,	4.,	.50,	29.2e5,	'Evergreen Needleleaf Forest'	
15,	15.,	.26,	.94,	69.,	4.,	.54,	41.8e5,	'Mixed Forest'	
16,	8.,	1.0,	.98,	0.01,	6.,	0.,	9.0e25,	'Water Bodies'	
17,	15.,	.60,	.95,	31.,	6.,	.55,	29.2e5,	'Herbaceous Wetland'	
18,	23.,	.51,	.95,	34.,	5.,	.58,	41.8e5,	'Wooded Wetland'	
19,	29.,	.03,	.85,	8.,	2.,	.62,	12.0e5,	'Barren or Sparsely Vegetated'	
20,	28.,	.41,	.92,	7.,	5.,	.60,	9.0e25,	'Herbaceous Tundra'	
21,	14.,	.61,	.93,	25.,	5.,	.60,	9.0e25,	'Wooded Tundra'	
22,	23.,	.43,	.92,	16.,	5.,	.60,	9.0e25,	'Mixed Tundra'	
23,	39.,	.05,	.85,	10.,	2.,	.62,	12.0e5,	'Bare Ground Tundra'	
24,	53.,	.93,	.95,	7.,	5.,	0.,	9.0e25,	'Snow or Ice'	
WINTER									
1,	15.,	.10,	.88,	110.,	3.,	.52,	18.9e5,	'Urban and Built-Up Land'	
2,	28.,	.66,	.92,	16.,	4.,	.60,	25.0e5,	'Dryland Cropland and Pasture'	
3,	21.,	.49,	.92,	4.,	4.,	.60,	25.0e5,	'Irrigated Cropland and Pasture'	
4,	22.,	.46,	.92,	4.,	4.,	.60,	25.0e5,	'Mixed Dryland/Irrigated Cropland and Pasture'	
5,	18.,	.21,	.92,	5.,	4.,	.60,	25.0e5,	'Cropland/Grassland Mosaic'	
6,	21.,	.39,	.93,	26.,	4.,	.60,	25.0e5,	'Cropland/Woodland Mosaic'	
7,	27.,	.30,	.92,	10.,	4.,	.60,	20.8e5,	'Grassland'	
8,	24.,	.21,	.88,	8.,	4.,	.62,	20.8e5,	'Shrubland'	
9,	24.,	.20,	.90,	10.,	4.,	.60,	20.8e5,	'Mixed Shrubland/Grassland'	
10,	21.,	.16,	.92,	23.,	3.,	0.,	25.0e5,	'Savanna'	
11,	15.,	.55,	.93,	60.,	5.,	.56,	25.0e5,	'Deciduous Broadleaf Forest'	
12,	16.,	.53,	.93,	103.,	5.,	.50,	25.0e5,	'Deciduous Needleleaf Forest'	
13,	11.,	.52,	.95,	280.,	5.,	0.,	29.2e5,	'Evergreen Broadleaf Forest'	
14,	10.,	.58,	.95,	51.,	5.,	.50,	29.2e5,	'Evergreen Needleleaf Forest'	
15,	16.,	.54,	.94,	45.,	6.,	.58,	41.8e5,	'Mixed Forest'	
16,	8.,	1.0,	.98,	0.01,	6.,	0.,	9.0e25,	'Water Bodies'	
17,	14.,	.80,	.95,	22.,	6.,	.55,	29.2e5,	'Herbaceous Wetland'	
18,	17.,	.59,	.95,	34.,	6.,	.58,	41.8e5,	'Wooded Wetland'	
19,	28.,	.05,	.85,	10.,	2.,	.62,	12.0e5,	'Barren or Sparsely Vegetated'	
20,	36.,	.88,	.92,	6.,	5.,	0.,	9.0e25,	'Herbaceous Tundra'	
21,	42.,	.91,	.93,	45.,	5.,	0.,	9.0e25,	'Wooded Tundra'	
22,	53.,	.88,	.92,	22.,	5.,	0.,	9.0e25,	'Mixed Tundra'	
23,	57.,	.89,	.95,	6.,	5.,	0.,	12.0e5,	'Bare Ground Tundra'	
24,	68.,	.96,	.95,	4.,	5.,	0.,	9.0e25,	'Snow or Ice'	

Table 15 continued:

## LANDUSE.TBL.plus05

USGS	24,2	'ALBD	SLMO	SPFM	SFZO	THERIN	SCFX	SFHC	
SUMMER									
1,	21.,	.08,	.88,	97.,	3.,	.52,	18.9e5,	'Urban and Built-Up Land'	
2,	15.,	.31,	.92,	19.,	4.,	.60,	25.0e5,	'Dryland Cropland and Pasture'	
3,	18.,	.71,	.92,	13.,	4.,	.60,	25.0e5,	'Irrigated Cropland and Pasture'	
4,	16.,	.28,	.92,	24.,	4.,	.60,	25.0e5,	'Mixed Dryland/Irrigated Cropland and Pasture'	
5,	16.,	.26,	.92,	12.,	4.,	.60,	25.0e5,	'Cropland/Grassland Mosaic'	
6,	18.,	.40,	.93,	22.,	4.,	.60,	25.0e5,	'Cropland/Woodland Mosaic'	
7,	14.,	.18,	.92,	11.,	3.,	.60,	20.8e5,	'Grassland'	
8,	21.,	.13,	.88,	14.,	3.,	.62,	20.8e5,	'Shrubland'	
9,	13.,	.13,	.90,	13.,	3.,	.60,	20.8e5,	'Mixed Shrubland/Grassland'	
10,	23.,	.18,	.92,	15.,	3.,	0.,	25.0e5,	'Savanna'	
11,	23.,	.30,	.93,	85.,	4.,	.56,	25.0e5,	'Deciduous Broadleaf Forest'	
12,	12.,	.33,	.94,	83.,	4.,	.50,	25.0e5,	'Deciduous Needleleaf Forest'	
13,	14.,	.44,	.95,	90.,	5.,	0.,	29.2e5,	'Evergreen Broadleaf Forest'	
14,	13.,	.58,	.95,	49.,	4.,	.50,	29.2e5,	'Evergreen Needleleaf Forest'	
15,	10.,	.55,	.94,	64.,	4.,	.54,	41.8e5,	'Mixed Forest'	
16,	8.,	1.0,	.98,	0.01,	6.,	0.,	9.0e25,	'Water Bodies'	
17,	16.,	.71,	.95,	18.,	6.,	.55,	29.2e5,	'Herbaceous Wetland'	
18,	16.,	.58,	.95,	39.,	5.,	.58,	41.8e5,	'Wooded Wetland'	
19,	22.,	.05,	.85,	7.,	2.,	.62,	12.0e5,	'Barren or Sparsely Vegetated'	
20,	22.,	.45,	.92,	11.,	5.,	.60,	9.0e25,	'Herbaceous Tundra'	
21,	25.,	.74,	.93,	36.,	5.,	.60,	9.0e25,	'Wooded Tundra'	
22,	10.,	.50,	.92,	29.,	5.,	.60,	9.0e25,	'Mixed Tundra'	
23,	26.,	.13,	.85,	8.,	2.,	.62,	12.0e5,	'Bare Ground Tundra'	
24,	60.,	.94,	.95,	5.,	5.,	0.,	9.0e25,	'Snow or Ice'	
WINTER									
1,	16.,	.20,	.88,	59.,	3.,	.52,	18.9e5,	'Urban and Built-Up Land'	
2,	32.,	.63,	.92,	8.,	4.,	.60,	25.0e5,	'Dryland Cropland and Pasture'	
3,	16.,	.58,	.92,	7.,	4.,	.60,	25.0e5,	'Irrigated Cropland and Pasture'	
4,	23.,	.49,	.92,	6.,	4.,	.60,	25.0e5,	'Mixed Dryland/Irrigated Cropland and Pasture'	
5,	19.,	.37,	.92,	3.,	4.,	.60,	25.0e5,	'Cropland/Grassland Mosaic'	
6,	19.,	.57,	.93,	24.,	4.,	.60,	25.0e5,	'Cropland/Woodland Mosaic'	
7,	24.,	.36,	.92,	9.,	4.,	.60,	20.8e5,	'Grassland'	
8,	24.,	.18,	.88,	8.,	4.,	.62,	20.8e5,	'Shrubland'	
9,	22.,	.27,	.90,	12.,	4.,	.60,	20.8e5,	'Mixed Shrubland/Grassland'	
10,	24.,	.10,	.92,	14.,	3.,	0.,	25.0e5,	'Savanna'	
11,	21.,	.65,	.93,	42.,	5.,	.56,	25.0e5,	'Deciduous Broadleaf Forest'	
12,	17.,	.48,	.93,	57.,	5.,	.50,	25.0e5,	'Deciduous Needleleaf Forest'	
13,	15.,	.57,	.95,	71.,	5.,	0.,	29.2e5,	'Evergreen Broadleaf Forest'	
14,	17.,	.47,	.95,	70.,	5.,	.50,	29.2e5,	'Evergreen Needleleaf Forest'	
15,	22.,	.27,	.94,	71.,	6.,	.58,	41.8e5,	'Mixed Forest'	
16,	8.,	1.0,	.98,	0.01,	6.,	0.,	9.0e25,	'Water Bodies'	
17,	14.,	.73,	.95,	11.,	6.,	.55,	29.2e5,	'Herbaceous Wetland'	
18,	13.,	.50,	.95,	38.,	6.,	.58,	41.8e5,	'Wooded Wetland'	
19,	26.,	.04,	.85,	8.,	2.,	.62,	12.0e5,	'Barren or Sparsely Vegetated'	
20,	51.,	.78,	.92,	10.,	5.,	0.,	9.0e25,	'Herbaceous Tundra'	
21,	49.,	.95,	.93,	24.,	5.,	0.,	9.0e25,	'Wooded Tundra'	
22,	59.,	.81,	.92,	15.,	5.,	0.,	9.0e25,	'Mixed Tundra'	
23,	61.,	.93,	.95,	2.,	5.,	0.,	12.0e5,	'Bare Ground Tundra'	
24,	75.,	.95,	.95,	5.,	5.,	0.,	9.0e25,	'Snow or Ice'	

Table 15 continued:

LANDUSE.TBL.plus06

USGS									
24,2,	'ALBD	SLMO	SFEM	SFZ0	THERIN	SCFX	SFHC		
SUMMER									
1,	21.,	.17,	.88,	94.,	3.,	.52,	18.9e5,	'Urban and Built-Up Land'	
2,	14.,	.51,	.92,	17.,	4.,	.60,	25.0e5,	'Dryland Cropland and Pasture'	
3,	16.,	.50,	.92,	11.,	4.,	.60,	25.0e5,	'Irrigated Cropland and Pasture'	
4,	15.,	.37,	.92,	16.,	4.,	.60,	25.0e5,	'Mixed Dryland/Irrigated Cropland and Pasture'	
5,	18.,	.20,	.92,	11.,	4.,	.60,	25.0e5,	'Cropland/Grassland Mosaic'	
6,	21.,	.31,	.93,	26.,	4.,	.60,	25.0e5,	'Cropland/Woodland Mosaic'	
7,	18.,	.18,	.92,	8.,	3.,	.60,	20.8e5,	'Grassland'	
8,	23.,	.16,	.88,	16.,	3.,	.62,	20.8e5,	'Shrubland'	
9,	24.,	.13,	.90,	12.,	3.,	.60,	20.8e5,	'Mixed Shrubland/Grassland'	
10,	26.,	.13,	.92,	18.,	3.,	0.,	25.0e5,	'Savanna'	
11,	15.,	.32,	.93,	43.,	4.,	.56,	25.0e5,	'Deciduous Broadleaf Forest'	
12,	18.,	.27,	.94,	68.,	4.,	.50,	25.0e5,	'Deciduous Needleleaf Forest'	
13,	7.,	.36,	.95,	107.,	5.,	0.,	29.2e5,	'Evergreen Broadleaf Forest'	
14,	12.,	.37,	.95,	67.,	4.,	.50,	29.2e5,	'Evergreen Needleleaf Forest'	
15,	17.,	.47,	.94,	73.,	4.,	.54,	41.8e5,	'Mixed Forest'	
16,	8.,	1.0,	.98,	0.01,	6.,	0.,	9.0e25,	'Water Bodies'	
17,	15.,	.66,	.95,	12.,	6.,	.55,	29.2e5,	'Herbaceous Wetland'	
18,	19.,	.35,	.95,	51.,	5.,	.58,	41.8e5,	'Wooded Wetland'	
19,	23.,	.02,	.85,	12.,	2.,	.62,	12.0e5,	'Barren or Sparsely Vegetated'	
20,	30.,	.63,	.92,	8.,	5.,	.60,	9.0e25,	'Herbaceous Tundra'	
21,	18.,	.52,	.93,	32.,	5.,	.60,	9.0e25,	'Wooded Tundra'	
22,	34.,	.51,	.92,	14.,	5.,	.60,	9.0e25,	'Mixed Tundra'	
23,	23.,	.21,	.85,	9.,	2.,	.62,	12.0e5,	'Bare Ground Tundra'	
24,	54.,	.96,	.95,	4.,	5.,	0.,	9.0e25,	'Snow or Ice'	
WINTER									
1,	23.,	.07,	.88,	95.,	3.,	.52,	18.9e5,	'Urban and Built-Up Land'	
2,	22.,	.69,	.92,	8.,	4.,	.60,	25.0e5,	'Dryland Cropland and Pasture'	
3,	22.,	.52,	.92,	9.,	4.,	.60,	25.0e5,	'Irrigated Cropland and Pasture'	
4,	25.,	.51,	.92,	4.,	4.,	.60,	25.0e5,	'Mixed Dryland/Irrigated Cropland and Pasture'	
5,	23.,	.38,	.92,	5.,	4.,	.60,	25.0e5,	'Cropland/Grassland Mosaic'	
6,	26.,	.44,	.93,	16.,	4.,	.60,	25.0e5,	'Cropland/Woodland Mosaic'	
7,	19.,	.29,	.92,	4.,	4.,	.60,	20.8e5,	'Grassland'	
8,	21.,	.25,	.88,	10.,	4.,	.62,	20.8e5,	'Shrubland'	
9,	25.,	.28,	.90,	13.,	4.,	.60,	20.8e5,	'Mixed Shrubland/Grassland'	
10,	19.,	.13,	.92,	14.,	3.,	0.,	25.0e5,	'Savanna'	
11,	16.,	.65,	.93,	92.,	5.,	.56,	25.0e5,	'Deciduous Broadleaf Forest'	
12,	17.,	.56,	.93,	39.,	5.,	.50,	25.0e5,	'Deciduous Needleleaf Forest'	
13,	13.,	.57,	.95,	50.,	5.,	0.,	29.2e5,	'Evergreen Broadleaf Forest'	
14,	17.,	.61,	.95,	58.,	5.,	.50,	29.2e5,	'Evergreen Needleleaf Forest'	
15,	11.,	.50,	.94,	147.,	6.,	.58,	41.8e5,	'Mixed Forest'	
16,	8.,	1.0,	.98,	0.01,	6.,	0.,	9.0e25,	'Water Bodies'	
17,	15.,	.68,	.95,	12.,	6.,	.55,	29.2e5,	'Herbaceous Wetland'	
18,	13.,	.70,	.95,	32.,	6.,	.58,	41.8e5,	'Wooded Wetland'	
19,	26.,	.04,	.85,	6.,	2.,	.62,	12.0e5,	'Barren or Sparsely Vegetated'	
20,	55.,	.87,	.92,	9.,	5.,	0.,	9.0e25,	'Herbaceous Tundra'	
21,	45.,	.91,	.93,	47.,	5.,	0.,	9.0e25,	'Wooded Tundra'	
22,	35.,	.44,	.92,	17.,	5.,	0.,	9.0e25,	'Mixed Tundra'	
23,	72.,	.51,	.95,	17.,	5.,	0.,	12.0e5,	'Bare Ground Tundra'	
24,	69.,	.91,	.95,	7.,	5.,	0.,	9.0e25,	'Snow or Ice'	

Table 15 continued:

LANDUSE.TBL.plus07

USGS	24,2	'ALBD	SLMO	SFEM	SFZO	THERIN	SCFX	SFHC	
SUMMER									
1,	26.,	.12,	.88,	53.,	3.,	.52,	18.9e5,	'Urban and Built-Up Land'	
2,	18.,	.20,	.92,	23.,	4.,	.60,	25.0e5,	'Dryland Cropland and Pasture'	
3,	16.,	.41,	.92,	15.,	4.,	.60,	25.0e5,	'Irrigated Cropland and Pasture'	
4,	17.,	.20,	.92,	12.,	4.,	.60,	25.0e5,	'Mixed Dryland/Irrigated Cropland and Pasture'	
5,	17.,	.41,	.92,	12.,	4.,	.60,	25.0e5,	'Cropland/Grassland Mosaic'	
6,	25.,	.33,	.93,	27.,	4.,	.60,	25.0e5,	'Cropland/Woodland Mosaic'	
7,	19.,	.14,	.92,	12.,	3.,	.60,	20.8e5,	'Grassland'	
8,	23.,	.14,	.88,	6.,	3.,	.62,	20.8e5,	'Shrubland'	
9,	21.,	.24,	.90,	12.,	3.,	.60,	20.8e5,	'Mixed Shrubland/Grassland'	
10,	16.,	.12,	.92,	18.,	3.,	0.,	25.0e5,	'Savanna'	
11,	16.,	.34,	.93,	52.,	4.,	.56,	25.0e5,	'Deciduous Broadleaf Forest'	
12,	15.,	.32,	.94,	41.,	4.,	.50,	25.0e5,	'Deciduous Needleleaf Forest'	
13,	15.,	.55,	.95,	47.,	5.,	0.,	29.2e5,	'Evergreen Broadleaf Forest'	
14,	15.,	.35,	.95,	131.,	4.,	.50,	29.2e5,	'Evergreen Needleleaf Forest'	
15,	14.,	.26,	.94,	38.,	4.,	.54,	41.8e5,	'Mixed Forest'	
16,	8.,	1.0,	.98,	0.01,	6.,	0.,	9.0e25,	'Water Bodies'	
17,	14.,	.56,	.95,	29.,	6.,	.55,	29.2e5,	'Herbaceous Wetland'	
18,	16.,	.49,	.95,	61.,	5.,	.58,	41.8e5,	'Wooded Wetland'	
19,	22.,	.02,	.85,	10.,	2.,	.62,	12.0e5,	'Barren or Sparsely Vegetated'	
20,	28.,	.44,	.92,	10.,	5.,	.60,	9.0e25,	'Herbaceous Tundra'	
21,	20.,	.70,	.93,	32.,	5.,	.60,	9.0e25,	'Wooded Tundra'	
22,	28.,	.59,	.92,	15.,	5.,	.60,	9.0e25,	'Mixed Tundra'	
23,	27.,	.01,	.85,	16.,	2.,	.62,	12.0e5,	'Bare Ground Tundra'	
24,	46.,	.95,	.95,	4.,	5.,	0.,	9.0e25,	'Snow or Ice'	
WINTER									
1,	21.,	.09,	.88,	70.,	3.,	.52,	18.9e5,	'Urban and Built-Up Land'	
2,	18.,	.43,	.92,	17.,	4.,	.60,	25.0e5,	'Dryland Cropland and Pasture'	
3,	25.,	.59,	.92,	7.,	4.,	.60,	25.0e5,	'Irrigated Cropland and Pasture'	
4,	22.,	.42,	.92,	10.,	4.,	.60,	25.0e5,	'Mixed Dryland/Irrigated Cropland and Pasture'	
5,	19.,	.38,	.92,	12.,	4.,	.60,	25.0e5,	'Cropland/Grassland Mosaic'	
6,	18.,	.51,	.93,	16.,	4.,	.60,	25.0e5,	'Cropland/Woodland Mosaic'	
7,	21.,	.31,	.92,	13.,	4.,	.60,	20.8e5,	'Grassland'	
8,	28.,	.18,	.88,	15.,	4.,	.62,	20.8e5,	'Shrubland'	
9,	28.,	.22,	.90,	10.,	4.,	.60,	20.8e5,	'Mixed Shrubland/Grassland'	
10,	25.,	.20,	.92,	14.,	3.,	0.,	25.0e5,	'Savanna'	
11,	20.,	.57,	.93,	59.,	5.,	.56,	25.0e5,	'Deciduous Broadleaf Forest'	
12,	15.,	.66,	.93,	87.,	5.,	.50,	25.0e5,	'Deciduous Needleleaf Forest'	
13,	11.,	.42,	.95,	97.,	5.,	0.,	29.2e5,	'Evergreen Broadleaf Forest'	
14,	11.,	.42,	.95,	127.,	5.,	.50,	29.2e5,	'Evergreen Needleleaf Forest'	
15,	16.,	.59,	.94,	133.,	6.,	.58,	41.8e5,	'Mixed Forest'	
16,	8.,	1.0,	.98,	0.01,	6.,	0.,	9.0e25,	'Water Bodies'	
17,	17.,	.73,	.95,	19.,	6.,	.55,	29.2e5,	'Herbaceous Wetland'	
18,	16.,	.76,	.95,	55.,	6.,	.58,	41.8e5,	'Wooded Wetland'	
19,	26.,	.06,	.85,	12.,	2.,	.62,	12.0e5,	'Barren or Sparsely Vegetated'	
20,	54.,	.91,	.92,	15.,	5.,	0.,	9.0e25,	'Herbaceous Tundra'	
21,	32.,	.80,	.93,	20.,	5.,	0.,	9.0e25,	'Wooded Tundra'	
22,	31.,	.87,	.92,	25.,	5.,	0.,	9.0e25,	'Mixed Tundra'	
23,	67.,	.91,	.95,	5.,	5.,	0.,	12.0e5,	'Bare Ground Tundra'	
24,	80.,	.95,	.95,	5.,	5.,	0.,	9.0e25,	'Snow or Ice'	

Table 15 continued:

LANDUSE.TBL.plus08

USGS										
24,2,	'ALBD	SLMO	SFEM	SFZO	THERIN	SCFX	SFHC			
SUMMER										
1,	20.,	.10,	.88,	47.,	3.,	.52,	18.9e5,	'Urban and Built-Up Land'		
2,	22.,	.18,	.92,	12.,	4.,	.60,	25.0e5,	'Dryland Cropland and Pasture'		
3,	28.,	.40,	.92,	13.,	4.,	.60,	25.0e5,	'Irrigated Cropland and Pasture'		
4,	13.,	.34,	.92,	13.,	4.,	.60,	25.0e5,	'Mixed Dryland/Irrigated Cropland and Pasture'		
5,	16.,	.24,	.92,	12.,	4.,	.60,	25.0e5,	'Cropland/Grassland Mosaic'		
6,	17.,	.31,	.93,	15.,	4.,	.60,	25.0e5,	'Cropland/Woodland Mosaic'		
7,	13.,	.30,	.92,	11.,	3.,	.60,	20.8e5,	'Grassland'		
8,	27.,	.08,	.88,	10.,	3.,	.62,	20.8e5,	'Shrubland'		
9,	19.,	.30,	.90,	12.,	3.,	.60,	20.8e5,	'Mixed Shrubland/Grassland'		
10,	20.,	.22,	.92,	14.,	3.,	0.,	25.0e5,	'Savanna'		
11,	15.,	.32,	.93,	108.,	4.,	.56,	25.0e5,	'Deciduous Broadleaf Forest'		
12,	17.,	.33,	.94,	116.,	4.,	.50,	25.0e5,	'Deciduous Needleleaf Forest'		
13,	13.,	.59,	.95,	86.,	5.,	0.,	29.2e5,	'Evergreen Broadleaf Forest'		
14,	12.,	.57,	.95,	33.,	4.,	.50,	29.2e5,	'Evergreen Needleleaf Forest'		
15,	16.,	.28,	.94,	68.,	4.,	.54,	41.8e5,	'Mixed Forest'		
16,	8.,	1.0,	.98,	0.01,	6.,	0.,	9.0e25,	'Water Bodies'		
17,	10.,	.58,	.95,	19.,	6.,	.55,	29.2e5,	'Herbaceous Wetland'		
18,	12.,	.31,	.95,	48.,	5.,	.58,	41.8e5,	'Wooded Wetland'		
19,	28.,	.01,	.85,	8.,	2.,	.62,	12.0e5,	'Barren or Sparsely Vegetated'		
20,	19.,	.63,	.92,	11.,	5.,	.60,	9.0e25,	'Herbaceous Tundra'		
21,	20.,	.49,	.93,	35.,	5.,	.60,	9.0e25,	'Wooded Tundra'		
22,	25.,	.59,	.92,	20.,	5.,	.60,	9.0e25,	'Mixed Tundra'		
23,	31.,	.17,	.85,	13.,	2.,	.62,	12.0e5,	'Bare Ground Tundra'		
24,	52.,	.88,	.95,	3.,	5.,	0.,	9.0e25,	'Snow or Ice'		
WINTER										
1,	15.,	.09,	.88,	36.,	3.,	.52,	18.9e5,	'Urban and Built-Up Land'		
2,	22.,	.55,	.92,	4.,	4.,	.60,	25.0e5,	'Dryland Cropland and Pasture'		
3,	22.,	.44,	.92,	10.,	4.,	.60,	25.0e5,	'Irrigated Cropland and Pasture'		
4,	22.,	.42,	.92,	3.,	4.,	.60,	25.0e5,	'Mixed Dryland/Irrigated Cropland and Pasture'		
5,	26.,	.26,	.92,	5.,	4.,	.60,	25.0e5,	'Cropland/Grassland Mosaic'		
6,	18.,	.57,	.93,	18.,	4.,	.60,	25.0e5,	'Cropland/Woodland Mosaic'		
7,	24.,	.24,	.92,	11.,	4.,	.60,	20.8e5,	'Grassland'		
8,	22.,	.21,	.88,	13.,	4.,	.62,	20.8e5,	'Shrubland'		
9,	28.,	.23,	.90,	12.,	4.,	.60,	20.8e5,	'Mixed Shrubland/Grassland'		
10,	21.,	.18,	.92,	14.,	3.,	0.,	25.0e5,	'Savanna'		
11,	13.,	.36,	.93,	51.,	5.,	.56,	25.0e5,	'Deciduous Broadleaf Forest'		
12,	17.,	.60,	.93,	66.,	5.,	.50,	25.0e5,	'Deciduous Needleleaf Forest'		
13,	11.,	.40,	.95,	58.,	5.,	0.,	29.2e5,	'Evergreen Broadleaf Forest'		
14,	10.,	.47,	.95,	84.,	5.,	.50,	29.2e5,	'Evergreen Needleleaf Forest'		
15,	13.,	.50,	.94,	95.,	6.,	.58,	41.8e5,	'Mixed Forest'		
16,	8.,	1.0,	.98,	0.01,	6.,	0.,	9.0e25,	'Water Bodies'		
17,	16.,	.67,	.95,	16.,	6.,	.55,	29.2e5,	'Herbaceous Wetland'		
18,	13.,	.67,	.95,	30.,	6.,	.58,	41.8e5,	'Wooded Wetland'		
19,	25.,	.03,	.85,	10.,	2.,	.62,	12.0e5,	'Barren or Sparsely Vegetated'		
20,	63.,	.54,	.92,	8.,	5.,	0.,	9.0e25,	'Herbaceous Tundra'		
21,	51.,	.78,	.93,	26.,	5.,	0.,	9.0e25,	'Wooded Tundra'		
22,	31.,	.90,	.92,	14.,	5.,	0.,	9.0e25,	'Mixed Tundra'		
23,	73.,	.82,	.95,	7.,	5.,	0.,	12.0e5,	'Bare Ground Tundra'		
24,	73.,	.89,	.95,	5.,	5.,	0.,	9.0e25,	'Snow or Ice'		



### References

- Anderson, J. L., 1996: A method of producing probabilistic forecasts from ensemble model integrations. *J. Climate*, **9**, 1518-1530.
- Baumhefner, D. P., National Center for Atmospheric Research, Boulder CO. Personal correspondence. July 20, 2000.
- Benjamin, S. G., J. M. Brown, K. J. Brundage, D. Dévényi, G. A. Grell, D. Kim, B. E. Schwartz, T. G. Smirnova, T. L. Smith, S. S. Weygandt, and G. S. Manikin, 2002: *RUC20 - The 20-km version of the Rapid Update Cycle*. <http://www.nws.noaa.gov/om/tpb/490body.htm>
- Bjerknes, V., T. Hesselberg, and O. Devik, 1911: *Dynamic meteorology and hydrography, part II. kinematics*. Carnegie Institute of Washington, D.C., 175 pp.
- Bretherton C., 2002: *ATMS 547 Boundary Layer Meteorology*. University of Washington course ATMS 547 Class Notes.
- Buizza, R., 1995: Optimal perturbation time evolution and sensitivity of ensemble prediction to perturbation amplitude. *Quart. J. Roy. Meteor. Soc.*, **121**, 1705-1738.
- , 1997: Potential forecast skill of ensemble prediction and spread and skill distributions of the ECMWF ensemble prediction system. *Mon. Wea. Rev.*, **125**, 99-119.
- , and T. N. Palmer, 1998: Impact of ensemble size on ensemble prediction. *Mon. Wea. Rev.*, **126**, 2503-2518.
- , M. Miller, and T. N. Palmer, 1999: Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **125**, 2887-2908.
- Brooks, H. E., and C. A. Doswell III, 1993: New technology and numerical weather prediction - a wasted opportunity? *Weather*, **48**, 173-177.
- Clancy, R. M., and W. D. Sadler, 1992: The Fleet Numerical Oceanography Center Suite of oceanographic models and products. *Wea. Forecasting*, **7**, 307-327.
- Cummings, J. A., Naval Research Laboratory, Marine Meteorology Division, Monterey, CA. Personal correspondence. July 26, 2002.
- Devore, J. L., 1995: *Probability and statistics for engineering and the sciences*, 4<sup>th</sup> ed. Belmont, CA: Wadsworth Press, 743 pp.
- Downton, R. A., and R. S. Bell, 1988: Analysis and model dependencies in medium-range forecast. *Meteorol. Mag.*, **117**, 279-285.
- Du, J., S. L. Mullen, and F. Sanders, 1997: Short-range forecasting of quantitative precipitation. *Mon. Wea. Rev.*, **125**, 2427-2459.

- Du, J., and M. S. Tracton, 2001: Implementation of a real-time short-range ensemble forecasting system at NCEP: an update. Preprints, *9th Conf. on Mesoscale Processes*, Ft. Lauderdale, FL, Amer. Meteor. Soc., 355-356.
- Dudhia, J. 1993: A nonhydrostatic version of the Penn State-NCAR mesoscale model: Validation tests and simulation of an Atlantic cyclone and cold front. *Mon. Wea. Rev.*, **121**, 1493-1513.
- Ebert, E. E., 2001: Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Mon. Wea. Rev.*, **129**, 2461-2480.
- Eckel, F. A., 1998: "Calibrated probabilistic quantitative precipitation forecasts based on the MRF ensemble." *Masters Thesis*, Air Force Institute of Technology, Dayton, OH, 133 pp.
- , and M. K. Walters, 1998: Calibrated probabilistic quantitative precipitation forecasts based on the MRF ensemble. *Wea. Forecasting*, **13**, 1132-1147.
- Epstein, E. S., 1969: Stochastic dynamic prediction. *Tellus*, **21**, 739-759.
- Errico, R. M., and D. P. Baumhefner, 1987: Predictability experiments using a high-resolution limited-area model. *Mon. Wea. Rev.*, **115**, 488-504.
- Errico, R. M., R. Langland, and D. P. Baumhefner, 2002: The workshop in atmospheric predictability. *Bull. Amer. Meteor. Soc.*, **83**, 1341-1344.
- Evans, R. E., M. S. J. Harrison, R. J. Graham, and K. R. Mylne, 2000: Joint medium-range ensembles from the Met. Office and ECMWF systems. *Mon. Wea. Rev.*, **128**, 3104-3127.
- Garvert, M., University of Washington Atmospheric Science Department, Seattle WA. Personal correspondence. Aug 15, 2002.
- Garratt, J. R., 1992: *The atmospheric boundary layer*. Cambridge, UK: Cambridge University Press, 316 pp.
- Gilmour, I., L. A. Smith, and R. Buizza, 2001: Linear regime duration: Is 24 hours a long time in synoptic weather forecasting? *J. Atmos. Sci.*, **58**, 3525-3539.
- Grimit, E. P., and C. F. Mass, 2002: Initial results of a mesoscale short-range ensemble forecasting system over the Pacific Northwest. *Wea. Forecasting*, **17**, 192-205.
- Grell, G. A., J. Dudhia, and D. R. Stauffer, 1994: A description of the fifth-generation Penn State/NCAR mesoscale model (MM5). NCAR/TN-398+STR, 121 pp [ Available from MMM Division, NCAR, P.O. Box 3000, Boulder, CO 80307]
- Hamill, T. M. and S. J. Colucci, 1997: Verification of Eta-RSM short-range ensemble forecasts. *Mon. Wea. Rev.*, **125**, 1312-1327.

- , S. L. Mullen, C. Snyder, Z. Toth, and D. P. Baumhefner, 2000a: Ensemble forecasting in the short to medium range: report from a workshop. *Bull. Amer. Meteor. Soc.*, **81**, 2653–2664.
- , C. Snyder, and R. E. Morss, 2000b: A comparison of probabilistic forecasts from bred, singular vector, and perturbed observation ensembles. *Mon. Wea. Rev.*, **128**, 1835–1851.
- , and C. Snyder, 2000c: A hybrid Kalman filter-3D Variational Analysis Scheme. *Mon. Wea. Rev.*, **128**, 2905–2919.
- , 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.*, **129**, 550–560.
- Hansen, J. A., 2002: Accounting for model error in ensemble-based state estimation and forecasting. *Mon. Wea. Rev.*, **130**, 2373–2391.
- Harrison, M. S. J., T. N. Palmer, D. Richardson, and R. Buizza, 1999: Analysis and model dependencies in medium-range ensembles: two transplant case-studies. *Quart. J. Roy. Meteor. Soc.*, **125**, 2487–2515.
- Houtekamer, P. L., J. Derome, 1995: Methods for ensemble prediction. *Mon. Wea. Rev.*, **123**, 2181–2196.
- , L. Lefaivre, J. Derome, H. Ritchie, and H. L. Mitchell, 1996: A system simulation approach to ensemble prediction. *Mon. Wea. Rev.*, **124**, 1225–1242.
- Jolliffe, I. T., and D. B. Stephenson, 2003: *Forecast Verification; A Practitioner's Guide in Atmospheric Science*. West Sussex, England: John Wiley & Sons, Ltd., 240 pp.
- Leith, C. E., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.*, **102**, 409–418.
- Lorenz, E. N., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci.*, **20**, 130–141.
- , 1969: The predictability of a flow which possesses many scales of motion. *Tellus*, **21**, 289–307.
- , 1993: *The Essence of Chaos*. Seattle, WA: University of Washington Press, 227 pp.
- Marzban, C., 2003: A comment on the ROC curve and the area under it as performance measures. *Wea. Forecasting* in press
- McMurdie, L., and C. Mass, 2003: Major numerical forecast failures over the Northwest Pacific. *Wea. Forecasting* in pres
- Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliaigis, 1996. The ECMWF ensemble prediction system: methodology and validation. *Quart. J. Roy. Meteor. Soc.*, **122**, 73–119.
- Mullen, S. L., and D. P. Baumhefner, 1988: Sensitivity to numerical simulations of explosive oceanic cyclogenesis. *Mon. Wea. Rev.*, **116**, 2289–2329.

- , 1989: The impact of initial condition uncertainty on numerical simulations of large-scale explosive cyclogenesis. *Mon. Wea. Rev.*, **117**, 2800–2821.
- , 1994: Monte Carlo simulations of explosive cyclogenesis. *Mon. Wea. Rev.*, **122**, 1548–1567.
- Mullen, S. L., J. Du, and F. Sanders, 1999: The dependence of ensemble dispersion on analysis-forecast systems: implications to short-range ensemble forecasting of precipitation. *Mon. Wea. Rev.*, **127**, 1674–1686.
- Murphy, A. H., 1998: The early history of probability forecasts: some extensions and clarifications. *Wea. and Fcst.*, **13**, 5–14.
- Murphy, J. M., 1988: The impact of ensemble forecasts on predictability. *Quart. J. Roy. Meteor. Soc.*, **114**, 463–494.
- Mylne, K. R., R. E. Evans, and R. T. Clark, 2002: Multimodel multianalysis ensembles in quasi-operational medium-range forecasting. *Quart. J. Roy. Meteor. Soc.*, **128**, 361–384.
- National Center for Atmospheric Research, Mesoscale and Microscale Meteorology Division, January 2000: *PSU/NCAR Mesoscale Modeling System, Tutorial Class Notes and User's Guide: MM5 Modeling System Version 3*.
- Nutter, P. A., 2003: "Effects of Nesting Frequency and Lateral Boundary perturbations on the Dispersion of Limited-Area Ensemble Forecasts." Ph.D. Dissertation, University of Oklahoma School of Meteorology, Norman, OK, 156 pp.
- Palmer, T. N., R. Mureau, and F. Molteni, 1990: The monte carlo forecast. *Weather.*, **45**, 198–207.
- Pielke, R. A., 2002: *Mesoscale meteorological modeling*, 2<sup>nd</sup> ed. San Diego, CA: Academic Press, 676 pp.
- Poincare, H., 1912: *Science et Methode*. Paris: Flammarion. English translation by F. Maitland, 1914: *Science and Method*. Thomas Nelson and Sons, 288 pp.
- Richardson, D. S., 2001a: Ensembles using multiple models and analyses. *Quart. J. Roy. Meteor. Soc.*, **127**, 1847–1864.
- Richardson, D. S., 2001b: Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Quart. J. Roy. Meteor. Soc.*, **127**, 2473–2498.
- Smagorinsky, J., 1969: Problems and promises of deterministic extended range forecasting. *Bull. Amer. Meteor. Soc.*, **50**, 286–3197.

- Stanski, H. R., L. J. Wilson, and W. R. Burrows, 1989: *Survey of common verification methods in meteorology*. Research Report No. 89-5, Downsview, Ontario: Environment Canada Atmospheric Environment Service, 114 pp.
- Stensrud, D. J., H. E. Brooks, J. Du, M. S. Tracton, and E. Rogers, 1999: Using ensembles for short-range forecasting. *Mon. Wea. Rev.*, **127**, 433–446.
- , J. Bao, and T. T. Warner, 2000: Using initial conditions and model physics perturbations in short-range ensemble simulations of mesoscale convective systems. *Mon. Wea. Rev.*, **128**, 2077–2107.
- Tallagrad, O., R. Vautard, and B. Strauss, 1999: Evaluation of probabilistic prediction systems. In *Proceedings from the Workshop on Predictability*, 20–22 October 1997, European Center for Medium-Range Weather Forecasts. 1–25.
- Toth, Z. and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.*, **74**, 2317–2330.
- Tracton, M. S., and E. Kalnay, 1993: Operational ensemble prediction at the National Meteorological Center: practical aspects. *Wea. Forecasting*, **8**, 379–398.
- Wandishin, M. S., S. L. Mullen, D. J. Stensrud, and H. E. Brooks, 2001: Evaluation of a short-range multimodel ensemble system. *Mon. Wea. Rev.*, **129**, 729–747.
- Wilks, D. S., 1995: *Statistical methods in the atmospheric sciences: an introduction*. San Diego, CA: Academic Press, 467 pp.
- Ziehmann, C., 2000: Comparison of a single-model EPS with a multimodel ensemble consisting of a few operational models. *Tellus*, **52A**, 280–299.

Vita

Major Frederick Anthony Eckel was born on [REDACTED] in Albany, NY. He graduated from Bethlehem Central High School in 1985. In May 1989, he received a Bachelor of Science in Physics from the State University of New York at Cortland. Concurrently, he received a commission in the USAF having completed the Air Force Reserve Officer Training Corps program at Detachment 520, Cornell University. In March 1998, he received a Masters of Science in Atmospheric Sciences from the Air Force Institute of Technology (AFIT) where he was also honored with the Commandant's Award for the most outstanding Master's thesis in the entire school.

Major Eckel's first assignment after completing the 1-year Basic Meteorology Program at Texas A & M University was the Wing Weather Officer at McChord AFB, WA. As part of that position, he deployed as the tactical weather team leader to Cairo West Air Base, Egypt, in support of Operation Restore Hope. Next, he was assigned as Chief of Weather Operations at Yokota AFB, Japan, where he was recognized as an Exceptional Performer by the Air Weather Service Standardization and Evaluation team and also received the PACAF Weather Officer of the Year award. After attending AFIT in residence in Dayton, OH, he became the Air Force Weather Agency's (AFWA) liaison to the Space Warfare Center, Schriever AFB, CO. Following graduation, he will be assigned to HQ AFWA, Offutt AFB, NE, to lead the Meteorological Models Branch of the Air and Space Sciences Directorate.